Karlsruhe Institute of Technology

Pronominal Anaphora in Machine Translation

Master Thesis
of

Jochen Stefan Weiner

Institute for Anthropomatics and Robotics
Interactive Systems Lab (ISL)

Reviewer:	Prof. Dr. Alex Waibel
Second reviewer:	Dr. Sebastian Stüker
Advisors:	Dipl.-Inform. Jan Niehues
	Teresa Herrmann, M.Sc.

Duration: August 01, 2013 – January 31, 2014

Abstract

State-of-the-art machine translation systems use strong assumptions of independence. Following these assumptions language is split into small segments such as sentences and phrases which are translated independently. Natural language, however, is not independent: many concepts depend on context. One such case is reference introduced by pronominal anaphora. In pronominal anaphora a pronoun word (anaphor) refers to a concept mentioned earlier in the text (antecedent). This type of reference can refer to something in the same sentence, but it can also span many sentences. Pronominal anaphora pose a challenge for translators since the anaphor has to fulfil some grammatical agreement with the antecedent. This means that the reference has to be detected in the source text before translation and the translator needs to ensure that this reference still holds true in the translation. The independence assumptions of current machine translation systems do not allow for this.

We study pronominal anaphora in two tasks of English–German machine translation. We analyse occurrence of pronominal anaphora and their current translation performance. In this analysis we find that the implicit handling of pronominal anaphora in our baseline translation system is not sufficient. Therefore we develop four approaches to handle pronominal anaphora explicitly. Two of these approaches are based on post-processing. In the first one we correct pronouns directly and in the second one we select a hypothesis with correct pronouns from the translation system's n-best list. Both of these approaches improve the translation accuracy of the pronouns but hardly change the translation quality measured in BLEU. The other two approaches predict translations of pronoun words and can be used in the decoder. The Discriminative Word Lexicon (DWL) predicts the probability of a target word to be used in the translation and the Source DWL (SDWL) directly predicts the translation of a source language pronoun. However, these predictions do not improve the quality already achieved by the translation system.

Zusammenfassung

Bestehende maschinelle Übersetzungssysteme beruhen auf starken Unabhängigkeitsannahmen. Unter diesen Annahmen wird ein Eingabetext in kleine Einheiten wie Sätze oder Phrasen unterteilt, die dann unabhängig voneinander übersetzt werden. Natürliche Sprache besteht jedoch nicht aus unabhängigen Einheiten. Abhängigkeiten entstehen beispielsweise durch Anaphorik. Pronominale Anaphorik ist ein linguistisches Konzept, das Verbindungen von einem Pronomen (Anaphor) zu einem Konzept aufbaut, das bereits im Satz genannt worden ist (Antezedens). Diese Verbindung kann innerhalb eines Satzes bestehen, sie kann aber auch über mehrere Sätze hinweg gehen. Pronominale Anaphorik stellt eine Herausforderung für die Übersetzung dar, denn eine Anaphor ist dadurch gekennzeichnet, dass sie eine gewisse grammatische Übereinstimmung mit dem Antezedens aufweist. Das bedeutet, dass die Verbindung zwischen Anaphor und Antezedens vor der Übersetzung erkannt und dann richtig in die Zielsprache übertragen werden muss. Durch die starken Unabhängigkeitsannahmen aktueller maschineller Übersetzungssysteme ist ein solches Vorgehen für diese Systeme nicht möglich.

Wir untersuchen pronominale Anaphorik in zwei verschiedenen Textarten für Englisch–Deutsche Übersetzung. Wir analysieren das Auftreten von pronominaler Anaphorik und die Übersetzungsqualität unseres Übersetzungssystems. Die Analyse zeigt, dass das System die pronominale Anaphorik nur unzureichend gut übersetzt. Daher entwickeln wir vier Ansätze, die pronominale Anaphorik explizit betrachten. Zwei dieser Ansätze arbeiten mit fertigen Übersetzungshypothesen. Im ersten Ansatz werden Pronomen direkt korrigiert; im zweiten wird die Hypothese mit den meisten richtigen Pronomen aus der N-Besten-Liste ausgewählt. Diese Ansätze verbessern beide den Anteil der richtig übersetzten Pronomen, haben jedoch kaum Auswirkungen auf das BLEU Ergebnis. Die beiden anderen Ansätze schätzen die Übersetzung eines Pronomens und können im Decoder verwendet werden. Das Discriminative Word Lexicon (DWL) schätzt die Wahrscheinlichkeit, dass ein Zielwort in der Übersetzung verwendet wird, während das Source DWL (SDWL) die Übersetzung des Pronomens direkt schätzt. Allerdings verbessern diese Abschätzungen die bereits bestehende Übersetzungsqualität nicht.

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 27. Januar 2014

Jochen Weiner

Acknowledgements

I would like to thank Jan Niehues and Teresa Herrmann for their advice during this research. I am grateful for the discussions with them and their suggestions that led me to new ideas. Their guidance and experience helped me complete this thesis in good time.

I am also grateful for the experience in research that I have been given at the Interactive Systems Lab. I learned a lot writing papers with others and taking part in the IWSLT 2013.

Contents

Abstract i

1 Introduction 1
1.1 Overview . . . 3

2 Fundamentals 4
2.1 Statistical Machine Translation . . . 4
2.2 Discriminative Word Lexicon . . . 6
2.3 BLEU . . . 7

3 Anaphora 8
3.1 Anaphora and Antecedent . . . 8
3.2 Pronouns . . . 9
3.3 Translating Pronominal Anaphora . . . 10
3.4 Pronominal Anaphora in Machine Translation . . . 11

4 Related Work 13
4.1 Explicit Pronominal Anaphora Handling in MT . . . 13
4.1.1 Phrase-Based MT . . . 13
4.1.2 Deep Syntactic MT . . . 14
4.2 Integration of Other Connectives into MT . . . 15
4.3 Discourse-Level Translation . . . 15
4.4 Evaluating Pronoun Translation . . . 15

5 Resources 17
5.1 Translation Tasks . . . 17
5.2 Part-of-Speech Tags . . . 17
5.2.1 Part-of-Speech Taggers . . . 18
5.2.2 Finegrained POS Tags for German . . . 18
5.3 Anaphora Resolution . . . 19
5.4 Resolution Translation and Evaluation . . . 21
5.5 Sources of Error . . . 22

6 Analysing Pronominal Anaphora 23
6.1 Pronominal Anaphora in Text . . . 23
6.2 Intra-Sentential and Inter-Sentential Anaphora . . . 25
6.3 Translation of Source Pronouns . . . 26

7 Post-Processing Based On Anaphora Resolution **30**
7.1 Correcting Translations of Anaphors 30
7.2 Correcting Incorrect Pronouns 31
7.2.1 Changed Pronouns 31
7.2.2 BLEU Scores of Resulting Translation Text 36
7.2.3 Translation of Source Pronouns 37
7.3 N-Best Hypothesis Selection 39
7.3.1 Changed Pronouns 39
7.3.2 BLEU Scores of Resulting Translation Text 40
7.3.3 Translation of Source Pronouns 41

8 Discriminative Word Lexica for Pronouns **43**
8.1 Features for a Discriminative Word Lexicon 44
8.1.1 Extra Features 44
8.2 Evaluation for Pronouns 47

9 Source Discriminative Word Lexica for Pronouns **52**
9.1 Model Types 52
9.2 Features 53
9.3 Evaluation for Pronouns 54

10 Comparison of the Approaches **56**

11 Conclusion **60**
11.1 Outlook 61

Nomenclature **63**

Bibliography **65**

1. Introduction

Modern systems for statistical machine translation are already quite successful. Depending on the language pair they are able to produce reasonable or even good translations. However, these systems are limited by strong assumptions of locality or independence. Under these assumptions the text to be translated is split into many small units that are translated independently of one another. The strongest independence assumption states the full independence of sentences: state-of-the-art systems independently translate sentences one by one without regard to the sentences around them. The sentence-level independence assumption is not the only independence assumption. Many translation systems use phrase-based translation models. These translation models split the sentence into individual phrases which are hardly ever longer than a few words. This translation approach has the built-in assumption that phrases can be translated independently. Other models in the log-linear model such as the language model go beyond the phrase. However, the history of an n-gram language model typically does not cover more than three or four words and assumes the current words to be independent from everything before that. While the language model may be able to link individually translated phrases together, it is not able to model long range relationships. These assumptions are strong limitations for the translation system. For practical reasons translation systems ignore problems and phenomena that go beyond the phrase-level and thus make language coherent.

From a linguistic point of view these limitations are highly problematic since they do not reflect the nature of natural language. There are many different phenomena that introduce dependence within or across sentences and contradict the independence assumptions of the translation system. One such phenomenon is the reference to something mentioned earlier in the text:

(1) *When the girl went outside,* ***she*** *put on* ***her*** *hat.*
But ***she*** *could still feel the cold.*

(2) *When the bear felt winter was coming,* ***it*** *went into* ***its*** *den.*
There ***it*** *prepared for hibernation.*

This type of reference, called *pronominal anaphora*, is very common. In the first example the pronouns *she* and *her* refer back to the word *girl*, in the second example

it and *its* refer to *bear*. The referring word (the *anaphor*) does not have a meaning by itself, but depends on the word it refers to (the *antecedent*) for its interpretation. Therefore a translator needs to identify this reference and reflect it in the translation.

In most languages the reference between antecedent and anaphor is marked my some sort of grammatical agreement between these two words. When translating pronominal anaphora, the translator has to ensure that the translation of the anaphor correctly refers to the translation of the antecedent. Since there are often many different words into which a word can be translated, the translator needs to take into account how the antecedent was translated in order to ensure the anaphor correctly refers to it.

Given the independence assumptions employed by state-of-the-art machine translation systems, they have no way of identifying these pronouns and taking their reference into account. When the anaphoric reference goes beyond the sentence boundary, the translation system has no means of discovering this relationship. Whether or not the pronoun is translated correctly will completely be down to chance. For anaphoric reference within a sentence the translation systems are limited by the independence assumptions built into phrase-based translation models and language models. While there are cases in which the phrase-based model has a phrase translation with the correct pronoun translation, there are also cases in which this is not the case. In the same way the language model may have seen the correct pronoun translation, but it is also possible that it has not seen the correct pronoun translation. So whether or not the pronoun is translated correctly depends on the context seen in training and not on the actual antecedent. This is problematic because in most contexts it is linguistically possible to replace, for example, a male actor by a female actor. The translation system should produce translations for the two cases that only differ in the words that mark the different actors. Since the translation model can only build on what it has seen during training, it will not be able to distinguish this subtle but important difference. There is no way of knowing whether or not the translation system is capable of producing a correct translation.

In this thesis we study pronominal anaphora in English–German machine translation. We analyse occurrence and translation of pronominal anaphora on two different translation tasks. Furthermore, we investigate the changes necessary to ensure that all pronominal anaphora are translated correctly. We conduct these experiments to find out whether the implicit pronoun handling in our baseline translation system is already sufficient and what results we would achieve if all pronouns were translated correctly.

Following this analysis we develop four approaches to handling pronominal anaphora explicitly: two approaches post-process a given translation, while the other two influence the decoding procedure by predicting the correct translation of a pronoun.

1.1 Overview

The work on pronominal anaphora in machine translation presented in this thesis is structured as follows:

Chapter 2 "Fundamentals" introduces the basic principles of machine translation. In addition to these basics it gives a detailed description of the Discriminative Word Lexicon (DWL). The chapter closes with a description of the evaluation metric BLEU.

Chapter 3 "Anaphora" introduces the concept of anaphora. Since this thesis is about translating pronominal anaphora, we first give a description of the linguistic concept of anaphora before turning to factors that are important for the translation of anaphora and the difficulties machine translation systems face when translating anaphora.

Chapter 4 "Related Work" describes work related to handling anaphora resolution in machine translation.

Chapter 5 "Resources" gives an overview over the two translation tasks that we work with in this thesis. The chapter describes the data sources used and the tools used to obtain this data. It provides a detailed description of the the method we use to automatically resolve anaphora.

Chapter 6 "Analysing Pronominal Anaphora" analyses pronominal anaphora in our data. We compare an automatic and a manual method for resolving anaphora. We report occurrence of anaphora as well as translation performance for these anaphora in the baseline translation system.

Chapter 7 "Post-Processing Based On Anaphora Resolution" describes our first two approaches to explicit handling of anaphora in machine translation. We use a list of resolved anaphora to (a) correct incorrectly translated words directly and (b) find a hypothesis with correct pronouns in the n-best list.

Chapter 8 "Discriminative Word Lexica for Pronouns" reports our third approach in which we investigate Discriminative Word Lexicon models for explicit and implicit anaphora handing.

Chapter 9 "Source Discriminative Word Lexica for Pronouns" describes our fourth and last approach to anaphora handling in machine translation which directly predicts the translation of an anaphor from features of the source sentence.

Chapter 10 "Comparison of the Approaches" provides and overview and a discussion of the results we obtained with our four approaches to explicit anaphora handling in machine translation.

Chapter 11 "Conclusion" concludes the work presented in this thesis and gives an outlook.

2. Fundamentals

We introduce the terms and concepts used in this thesis. First we outline the fundamental concepts of statistical machine translation (SMT). For in-depth information please refer to literature, such as the book "Statistical Machine Translation" by Philipp Koehn [Koe10]. We continue with a description of the Discriminative Word Lexicon which can be used in SMT. Finally we introduce the machine translation metric BLEU.

2.1 Statistical Machine Translation

The problem of machine translation is to translate a sentence $\mathbf{f}$ in the source language into a sentence $\hat{\mathbf{e}}$ in the target language. In terms of machine learning this means finding the target language sentence $\mathbf{e} = e_1, ..., e_J$ that out of all possible target language sentences E is the most probable for the given source language sentence $\mathbf{f} = f_1, ..., f_I$. Using knowledge from information theory in the *noisy channel model* and Bayes' theorem this is represented in the fundamental equation of machine translation:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e} \in E} p(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e} \in E} p(\mathbf{f}|\mathbf{e}) \cdot p(\mathbf{e}) \tag{2.1}$$

This equation, which was proposed by Brown et al. [BPPM93], laid the foundations of statistical machine translation. With this equation, the translation process can be broken down into three parts: The *translation model* provides $p(\mathbf{f}|\mathbf{e})$, the *language model* provides $p(\mathbf{e})$ and the *decoder* finds the best translation $\hat{\mathbf{e}}$.

The translation model (TM) provides estimates how likely the target sentence is a translation of the source sentence. The first translation models using the fundamental equation 2.1 were proposed by Brown et al. [BPPM93] together with the fundamental equation itself. These models are word-by-word translation models that try to find the best alignment between the words in the source sentence and words in the possible target sentence. Brown et al. describe a series of five increasingly complex algorithms that are trained on bilingual corpora. Nowadays these models are known as the *IBM models*[1].

[1] Brown et al. were at IBM at the time they proposed these models.

For many language pairs, there is no strict word to word correspondence. A translation word by word is, therefore, either not possible or results in suboptimal translations. Most state-of-the-art translation systems use the phrase-based machine translation approach (PBMT) [KOM03]. In this approach, the source sentence is not translated word by word but on a phrase basis. The sentence is split into non-overlapping phrases that each contain a few words. Each phrase is then translated into a target language phrase and the resulting phrases are reordered. In this way the system can easily produce translations that contain a different number of words than the source sentence while capturing the meaning more accurately. Phrases are not linguistically motivated, but extracted automatically. The extracted phrase pairs are kept in a *phrase table* together with their probabilities and further information. Since only phrases that have occurred several times in the training data are used in the phrase table, the word order in the target language phrase is usually correct. Thus PBMT implicitly also models reordering within a phrase. Phrase-based models have been shown to perform significantly better than word-by-word translation models. An example is shown in Figure 2.1

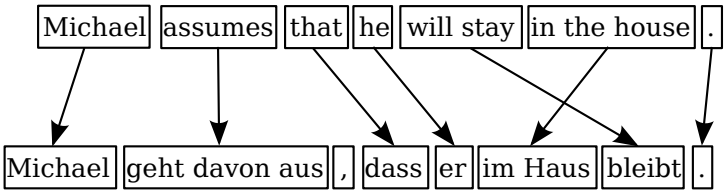


Figure 2.1: Phrase-based translation with reordering of phrases.

The language model (LM) provides an estimate how likely a sentence in the target language is a sentence of that language. A high LM score suggests that the sentence is a fluent and correct sentence. In many systems an n-gram language model is used. This model estimates the probability of a word given the history of the $n-1$ preceding words.

The decoder solves the search problem. From all possible word sequences in the target language it finds the one that is the best translation of the source sentence according to Equation 2.1.

In state-of-the-art SMT systems the noisy channel model (Equation 2.1) has been generalized into the *log-linear model*. This model is represented by the equation

$$\hat{\mathbf{e}} = \arg\min_{\mathbf{e}\in E} \sum_{i\in\mathcal{F}} -\lambda_i h_i(\mathbf{e}) \tag{2.2}$$

where $\mathcal{F}$ is a set of features, $h_i(\cdot)$ is a feature function and λ_i the weight for that feature. Equation 2.2 is equivalent to Equation 2.1 if we set

$$\begin{aligned} \mathcal{F} &= \{TM, LM\} \\ h_{TM}(\mathbf{e}) &= \log p(\mathbf{f}|\mathbf{e}) \\ h_{LM}(\mathbf{e}) &= \log p(\mathbf{e}) \end{aligned}$$

With the log-linear model the translation system is no longer restricted to translation model and language model. This modelling approach enables further models such as reordering model, phrase-count model, word-count model or discriminative word

this modelling approach to be included. Each of these models provides a feature function that returns a score from that model. This score is then weighted by the feature weight. The sum over all weighted feature scores is the score of the sentence $\mathbf{e}$. Through this simple model combination step each model can be trained and optimised individually. As a final training step, the weights need to be tuned, so that the influence of each model is set to such an amount that the models and weights together produce the best translations. This tuning is done with the *Minimum Error Rate Training* (MERT) [Och03]. As an instance of statistical machine learning SMT produces a number of hypotheses out of which it then chooses the best translation. The list with the n best translation hypotheses is called the n-best list. The MERT procedure tunes the model weights by iteratively adjusting them in such a way that in the resulting n-best list those hypotheses get better scores that are closer to a reference translation according to some metric such as BLEU (see Chapter 2.3).

2.2 Discriminative Word Lexicon

The *Discriminative Word Lexicon* (DWL) [BHK07, MHN09] is a model that uses features from the whole source sentence to predict the probability whether or not to include a target language word in the translation. The DWL is used as one model in the log-linear model approach and supports a fine-grained choice of words.

A *maximum entropy model* is trained to provide the probability of a target word given a set of features. In the original DWL model [MHN09] the words of the source sentence are used as features in the form of a bag-of-words. In the phrase-based translation approach models are often restricted to the current phrase, which means that phrases are translated independently of one another. The DWL, however, uses information from the whole sentence and can therefore model long range dependencies across phrases. Using a bag-of-words as features means that sentence structure is not taken into account. Sentence structure can be introduced to the model by adding additional features such as context on source and target side [NW13].

One binary maximum entropy classifier is trained for every target word. This classifier provides a probability whether or not the target word is to be included in the translation. Therefore positive and negative training examples must be created from the training data. Each training example contains a $label \in 0, 1$ marking it as a positive or negative example, and the set of features for that example.

positive examples

When the target word occurs in the reference translation of a sentence, we create a positive example [NW13].

negative examples

The naive approach is to create one negative example whenever the target word does not occur in the reference translation of a sentence. Since most words are only used in a few sentences, this would lead to highly unbalanced training examples [NW13].

In phrase-based translation, a translation is always based on phrase-pairs. A target word can only occur in the translation, if it appears in a target phrase for which the source phrase matches a part of the source sentence. We use the term *target vocabulary* to describe all these words that can occur in the

translation of a sentence. We create negative examples from sentences, for which the target word is in the target vocabulary but not in the reference translation [MCN+11, NW13]. This approach aims at achieving more balance between positive and negative examples and at reducing errors introduced by the phrase table.

The maximum entropy models trained on these training examples approximate the probability $p(e^+|feat_{\mathbf{f},e^+})$ of a target word $e^+ \in \mathbf{e}$ given the features $feat_{\mathbf{f},e^+}$ for source sentence $\mathbf{f} = f_1...f_I$ in combination with word e^+. The symbols e^+ and e^- denote the events that e is included or not included in the target sentence, respectively. Mauser et al. [MHN09] calculate this probability in the following way:

$$p(e^+|feat_{\mathbf{f},e^+}) = \frac{\exp\left(\sum_{f \in feat_{\mathbf{f},e^+}} \lambda_{f,e^+}\phi(f, feat_{\mathbf{f},e^+})\right)}{\sum_{e \in \{e^+,e^-\}} \exp\left(\sum_{f \in feat_{\mathbf{f},e}} \lambda_{f,e}\phi(f, feat_{\mathbf{f},e})\right)} \tag{2.3}$$

In this equation the $\lambda_{f,\cdot}$ are the feature weights and $\phi(f, feat_{\mathbf{f},e^+})$ are the simple feature functions

$$\phi(f, feat_{\mathbf{f},e^+}) = \begin{cases} 1 & \text{if} f \in feat_{\mathbf{f},e^+} \\ 0 & \text{else} \end{cases} \tag{2.4}$$

Using these probabilities for target words the probability for the target sentence $\mathbf{e} = e_1...e_J$ is then estimated as

$$p(\mathbf{e}|\mathbf{f}) = \prod_{e \in \mathbf{e}} p(e|feat_{\mathbf{f},e})$$

2.3 BLEU

BLEU, the *Bilingual Evaluation Understudy*, is an automatic evaluation metric for MT. It compares the translation output with the reference and looks for exact matches of words. The metric accounts for translation adequacy by including a word precision and translation fluency by including n-gram precision for 1-, 2-, 3- and 4-grams. It does not include recall, but instead has a brevity penalty that penalises very short translations. The final BLEU score is a weighted geometric average of the n-gram precisions p_n normalized with the brevity penalty BP:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right) \tag{2.5}$$

Usually there are a number of ways to translate a sentence. BLEU can use multiple references to account for this variability, but it does not account for synonyms or meaning. It does, therefore, not reflect small differences that make a huge impact in the meaning of a sentence.

3. Anaphora

3.1 Anaphora and Antecedent

Anaphora are linguistic elements that refer to some other linguistic element mentioned earlier in the same text [Cry04, Har12, TB01]. The linguistic element referred to by anaphora is called the *antecedent* [MCS95], and by definition anaphora depend on the antecedent for their interpretation [vDK00]. Anaphora allow recalling concepts that have already been introduced (represented by the antecedent) [BM00] without having to repeat these concepts again. As a very common phenomenon, anaphora occur in almost all types of text [HTS+11].

Anaphora may occur in two different contexts: they may either refer to an antecedent in the same sentence (intra-sentential anaphora) or to an antecedent in a previous sentence (inter-sentential anaphora) [LW03]. In the case of inter-sentential anaphora, the antecedent usually occurs within the n sentences preceding the anaphor, where n is close to one [KL75, Hob78].

There are several different types of anaphora which can involve pronouns, demonstrative determiners, pronominal substitution, ellipsis, verb-phrase and others [BM00]. This work concentrates on *pronominal anaphora* which is the type of anaphora in which the anaphor is a pronoun (see Chapter 3.2).

In order to understand, use and translate anaphora, the reference between anaphor and antecedent has to be identified. Only if the reader can correctly identify the concept a pronoun refers to, he can understand the text. Luckily, as humans, we are "amazingly good" [Nic03] at this task. In literature two different terms exist for this process of identifying reference in text: *coreference resolution* and *anaphora resolution*. The former refers to the process of "determining whether two expressions in natural language refer to the same entity in the world" [SNL01], regardless of their linguistic relationship in the text. The result is a *coreference chain* containing all the entities in the text referring to the same real world entity. *Anaphora resolution* on the other hand depends on linguistic relationships. This term describes the process of identifying anaphors and determining which linguistic entity in the text an anaphor refers to. It involves identifying the correct antecedent for anaphora, establishing

a connection between the two entities and merging previous information with the information supported by the anaphor [DMR83, Nic03]. While the terms coreference resolution and anaphora resolution in general describe completely distinct tasks[1], they may be used synonymously in the context of pronominal anaphora [LK10].

The term anaphora does not include linguistic elements referring forward to concepts occurring later in the text. These are called *cataphora* [Cry04].

3.2 Pronouns

A *pronoun*, grammatically speaking, is a word that stands for a noun, a noun-phrase or several noun phrases [Cry04, p. 210]. In terms of anaphora and antecedent, pronouns are those anaphora that are substituted by their antecedent noun phrase [LW03]. In the following example sentence, the word *it* is a pronoun anaphorically referring to its antecedent *apple*:

> *The girl took the apple and ate* ***it****.*

Pronouns are divided into several subclasses depending on the meaning they express. The following three subclasses [LW03, Cry04] are the so-called "central pronouns" [Cry04, p. 210] in the English language:

personal pronouns identify persons
nominative: *I, you, he, she, it, we, they*
objective: *me, you, him, her, it, us, them*

reflexive pronouns reflect the meaning of a noun phrase elsewhere
myself, yourself, himself, herself, itself, ourselves, yourselves, themselves

possessive pronouns express ownership
as determiners: *my, your, his, her, its, our, their*
on their own: *mine, yours, his, hers, its, ours, theirs*

Besides these, several other subclasses such as reciprocal, interrogative, relative, demonstrative, indefinite pronouns exist.

Some pronouns occur without any antecedent at all. These pronouns are called *pleonastic* or *structural* [LK10]. They are used when the syntax requires a pronoun, even if there is no antecedent for it to refer to. Examples include cases of the German *es* and the English *it*, as in the following sentence:

> *The girl went inside because* ***it*** *was raining.*

Here, the pronoun *it* does not refer to any linguistic entity mentioned earlier in the text but to the general concept of weather. Therefore this pronoun has no antecedent: it is used pleonastically.

In order to establish a connection between pronoun and antecedent, many languages demand some sort of grammatical agreement between pronoun and antecedent. Across languages, this demand ranges from relatively simple agreement to rather complex patterns of agreement [HF10].

[1] See [MEO+12] and [vDK00] for a detailed distinction of the two.

In the English language for example, some but not all pronouns require agreement in person, number and gender with their antecedent [Cry04]. In German, every pronoun also needs to agree with its antecedent in person, number and gender; but some cases also require agreement in politeness [Har12]. Other factors requiring agreement in some languages include humanness, animate/inanimate and emphasis.

3.3 Translating Pronominal Anaphora

When translating pronominal anaphora, it is important that the reference between pronoun and antecedent still holds true in the target language. However, the demands for agreement between anaphor and antecedent can vary strongly between languages (Chapter 3.2): the source language may require very different agreement patterns than the target language. This means that for most language pairs there is no one-to-one correspondence between pronouns. Indeed, for some pronouns the reference is very clear in one language but highly ambiguous in another [HF10]. The German *sie* is a personal pronoun which can either be feminine singular (to be translated as *she*, *her* or *it*), plural of all genders (*they* or *them*) or, capitalised, the polite form of address (second person singular and plural, *you*). In the other translation direction, the English pronoun *it* is translated into German as one of *er*, *sie* or *es*. Although English and German have similar agreement requirements (person, number, gender), there is no one-to-one correspondence between pronouns. These two languages use grammatical gender in different ways: While *it* can, when used anaphorically, refer to almost any noun phrase [NNZ13], the German pronoun depends on the grammatical gender of the noun.

(a) *The monkey ate the banana because* ***it*** *was hungry.*
(b) *The monkey ate the banana because* ***it*** *was ripe.*
(c) *The monkey ate the banana because* ***it*** *was tea-time.*

Example 1: Ambiguity of the word *it* [HS92].

The three sentences in Example 1 illustrate the difficulty of translating the word *it*. In all three cases the word *it* is a pronominal anaphor, but each time it refers to a different antecedent. In (a) the antecedent is *the monkey*. The word *monkey* translates into German as *Affe* which has masculine grammatical gender. Therefore the correct German translation of *it* in this sentence is the masculine German personal pronoun *er*. In (b) *it* refers to *the banana* which translates to the grammatically feminine word *Banane*. So in (b), *it* has to be translated as *sie*. In (c) the word *it* refers to the "abstract notion of time" [MCS95] and not to an entity earlier in the text. Since this is a pleonastic use of the pronoun (Chapter 3.2), *it* does not have an antecedent. The corresponding German pronoun for such pleonastic uses is *es*. In these three examples the word *it* has three different translations. If an incorrect pronoun is chosen in the translation, the translation would make no sense to the readers, leaving them mislead or confused [Gui12].

If the baby does not thrive on raw milk, boil ***it****.*

Example 2: Ambiguity with consequences [Jes54].

Example 2 shows a sentence where the pronoun is ambiguous. An incorrect choice of antecedent has severe consequences for the meaning of the translated sentence.

According to the English agreement patterns the anaphor *it* could refer to both *baby* and *milk*. In both cases the sentence would be grammatically correct. It is only the intention of the sentence that makes clear that the word *it* refers to *milk*. In German the sentence does not have this ambiguity: *Baby*, the translation of the English *baby*, has neutral grammatical gender. The pronoun *es* is used to refer to it. *Milk* on the the other hand translates as *Milch* which has feminine grammatical gender and thus requires the pronoun *sie*. If the antecedent *milk* is identified correctly, then *it* is correctly translated as *sie*. The translation correctly instructs to *boil the milk*. If, on the other hand, the naive translation *es* is chosen, the translation contains an incorrect reference to *baby*. The resulting sentence would instruct to *boil the baby*; a big error in the meaning of the sentence. If these incorrectly translated instructions were followed this could have severe consequences for the baby.

The translation difficulty in both cases derives from the fact that the anaphor itself does not contain a clue to which antecedent it refers to. The anaphor word itself is not enough to find the correct translation. Instead, the correct translation can only be created if the context is interpreted and the correct antecedent found. This shows that resolution of anaphora is of "crucial importance" [MCS95] for correct translation.

3.4 Pronominal Anaphora in Machine Translation

State-of-the-art phrase based machine translation systems are limited when it comes to translating pronominal anaphora. They assume sentences to be independent, and therefore translate them without regard to either their preceding or their following sentences [Har12]. In phrase-based translation a sentence is broken down into phrases. These phrases are hardly ever longer than a few words and translated independently of one another. This means the phrase based models assume that a sentence is made up of many small independent segments. Language Models and other models in the log-linear model soften the assumption of independence between individual phrases but are not able to overcome it. For reasons of practicality the history of an n-gram Language Model is hardly ever longer than three or four words. So while softening the independence between phrases, it does not introduce a large context. These factors contribute to an overall strong assumption of independence in MT.

Anaphora, on the other hand, introduce reference that links different elements in text together. If we only needed to know the source language antecedent in order to translate the anaphor, we could simply annotate the anaphor with its antecedent and then translate accordingly. Unfortunately, the problem is not as easy. The anaphor needs to agree with the antecedent grammatically, so its translation does not depend on the source language antecedent but on the antecedent translation. Therefore any model that assumes independence between these elements cannot reflect this reference: A given (antecedent) word can usually be translated into several different words in the target language. The anaphor needs to agree with the word actually chosen as a translation for the antecedent, so the translation system needs to determine the word that was chosen as a translation for the antecedent. Only then can it translate the anaphor properly [LW03, HF10, HTS+11].

For the translation of intra-sentential anaphora MT systems rely on the short history of the local Language Model (LM) and the context captured in phrases

[HF10, HTS+11]. This may lead to inconsistencies when the anaphor refers to an antecedent further away than the distance covered by LM history or phrases [HF10]. In Example 1 the distance between antecedent and anaphor in sentence (a) is five words, in (b) the distance is two words, and no distance can be defined for (c). The models may cover the distance of two words from *banana* to *it* in (b) either with a phrase or more probably with an n-gram in the language model; and there may be a phrase for *it was tea-time*. But the distance of five words from *monkey* to *it* in (a) is longer than a usual phrase and the history of a language model. Therefore it is too far for the models to implicitly reflect the reference. If then the pronoun in question and its context are ambiguous, the translation result will be essentially random [HF10].

For inter-sentential anaphora the problem goes further. The strict assumption of independence between sentences means that if there is a sentence boundary between antecedent and anaphor, none of the models will be able to reflect this reference, even if the distance between antecedent and anaphor is short. The system will be unable to determine the translation of the antecedent and can, therefore, not ensure it will chose an anaphor matching the antecedent. Instead the translation of the anaphor will only depend on local phrases [Gui12] and agreement with the antecedent will be down to chance [HTS+11].

I have a tree. ***It*** *is green.*

Example 3: Inter-sentential anaphora.

In the sentence pair in Example 3 the word *it* refers back to the word *tree* in the previous sentence. In English–German translation the correct translation of *tree* is *Baum* which has masculine grammatical gender. The correct translation of *it* would therefore be *er*. If the sentences are translated independently, the system will not be able to use this reference in the translation of *it*. Instead it will either translate this word according to the phrase *it is green* (if this phrase exists) or it will use the word *es* which is the naive translation of *it*.

These factors contribute to the conclusion that anaphora need to be handled explicitly in machine translation, if the system is to ensure they are translated correctly.

Even if there were a model that handles anaphora explicitly, the general performance of state-of-the-art SMT systems would still be a problem for handling anaphora [Har12]: A model supporting a small detail such as pronouns will not be able to do well, if the underlying baseline SMT system does not achieve a reasonably good translation result. If problems of word order or morphology are not resolved properly, it will not be possible to work on pronouns. Insufficient baseline performance has been reported to be problematic for a number of approaches for anaphora handling in machine translation ([HF10, Gui12], see Chapter 4.1.1). This leads Hardmeier to the conclusion that "there is little that researchers interested in anaphora can do about this problem except working on an easier language pair while waiting for the progress of SMT research in general" [Har12, p. 15].

4. Related Work

4.1 Explicit Pronominal Anaphora Handling in MT

There is little literature about explicit anaphora handling in machine translation. In the 1990's there was some research in connection with Rule-Based Machine Translation (RBMT). Since then the paradigm has moved away from RBMT. While the knowledge about the problem itself is still useful, those approaches to solving it are not applicable to modern MT systems [HF10].

Starting in 2010 the field has begun to attract attention again. Approaches have been proposed for phrase-based MT and for deep syntactic MT.

4.1.1 Phrase-Based MT

The approaches of Le Nagard and Koehn [LK10] and Hardmeier et al. [HF10] first employ a source language anaphora resolution tool in order to find anaphora and their antecedents in the text. They then decode a baseline translation and extract number and gender of the translation of the antecedents. This information is then used in two different ways:

Translating English to French, Le Nagard and Koehn only consider the pronouns *it* and *they* [LK10]. They only use the gender of the translated antecedent and annotate the anaphora on the source side with that gender. With this they introduce target language information into the source language input text. For example, the English word *it* is annotated to become *it-feminine* if the French reference translation of the antecedent is feminine. Number and case as additional agreement features are disregarded because there were too few occurrences of the different types in the corpus and the authors had problems with unreliable detection algorithms. Using this annotated text as their input, they re-train their SMT system and decode as usual. They report unchanged BLEU scores and a hardly improved number of correctly translated pronouns. They blame this on the poor performance of their anaphora resolution systems. Guillou employed the same approach for English to Czech translation [Gui12]. But instead of using anaphora resolution tools, she used

manually annotated anaphora resolution data. Despite this change towards good anaphora resolution, no real improvement is reported.

Translating English to German, Hardmeier et al. pair number and gender information of antecedents with their referring anaphor [HF10]. These pairs then act as the input for a new Word Dependency Model that acts as a feature function in a phrase-based SMT system. When the anaphor is translated, the system adds a score into the decoding process. They also report an unchanged BLEU score, but a small improvement in anaphor translation quality. Applying this same approach to the English to French translation task did not yield any improvements [HTS+11].

Although being two different approaches, these two methods share a number of problems. They both lead to pronoun over-generation, potentially because they favour pronouns as translations for source language pronouns which may not always be the adequate translation. Both approaches also suffer from insufficient performance of their anaphora resolution and antecedent translation spotting algorithms. In conclusion, neither of the two approaches has proven itself to be working accurately. They both need more "refinement before they can deliver consistently useful results" [Har12, p. 21]

The two approaches described above only use the connection between anaphor and its antecedent. Novák [Nov11] proposes the use of longer coreference chains that would enable a more confident translation choice, but no results on this proposal have been reported.

Popescu-Belis et al. [PBML+12] criticise two things in the annotation used in the two above approaches: First, the gender of the translated antecedent depends on the translation choice and is not fixed beforehand. Therefore the pronoun cannot a priori be annotated for certain. Second, depending on the language pair, other factors in addition to gender need to be taken into account. In order to avoid this and also to circumvent the errors introduced by anaphora resolution, they propose an approach in which pronouns are annotated without the need of anaphora resolution. Instead they employ human annotators to annotate pronouns in training data with their exact translation and then learn a model to do this automatically ("translation spotting"). They note that this does not avoid their above criticism that the pronoun translation cannot be determined a priori, but state that in their case of English to French translation this approach can work because of a very narrow range of possible translations. In fact, in their experiments, all correct translations of antecedents had the same gender as the reference. This implies that in their context the translation spotting method may be applicable, and in fact, they report a "small but significant" improvement of the translation's BLEU evaluation.

4.1.2 Deep Syntactic MT

Novák [Nov11] proposes several approaches for the integration of anaphora resolution into an MT system using deep syntactic (tectogrammatical) tree-to-tree transfer. Utilizing anaphora resolution on the source side, the pronoun's node in the tectogrammatical tree is annotated with the pronoun's antecedent, an approach conceptually similar to the two approaches cited above. In the tree-to-tree transfer's synthesis step gender and number are copied from the antecedent and the correct translation form is selected. In the special case of the translation of *it* from English to Czech,

this approach achieves some improvement in terms of correct translation of the pronoun [NNZ13]. Utilizing anaphora resolution on the target side, Novák proposes integrating resolution results into a tree language model in the hope for more reliable dependency relation estimates. No experimental results have been reported for this second proposal.

4.2 Integration of Other Connectives into MT

Meyer et al. present two methods for the integration of labels for discourse connectives [MPB12, MPBHG12]. Discourse connectives are words such as *although*, *however*, *since* or *while* that mark discourse relations between parts of texts. Unlike pronominal anaphora their translation depends on their sense and not on the actually chosen translation of another word (see Chapter 3.4). Therefore they do not depend on translation output, but can be annotated for certain before the translation process.

The first method modifies the phrase table [MPB12]. In this approach connectives are located in the phrase table and their sense in the translation determined. If the sense can be established, the phrase is changed by annotating the connective with that sense. With this they achieve some improvement in connective translation and a significant improvement in BLEU scores.

The second method [MPBHG12] uses Factored Translation Models [KH07]. From the connective source words and their sense labels they built feature vectors. These feature vectors could also include target language words but the authors state that this is not necessary for their task. With these feature vectors they train a Factored Translation Model and achieve small improvements in the number of correctly translated connectives but hardly any improvement in terms of BLEU scores.

4.3 Discourse-Level Translation

In order to overcome the limitations of the assumption that sentences can be handled individually (see Chapter 3.4), Stymne, Hardmeier et al. [HNT12, SHTN13] present a phrase-based translation algorithm that takes the whole discourse into account. Instead of the classical dynamic programming beam search algorithm on each sentence, they perform a hill climbing algorithm. The state of the hill climbing algorithm is a translation of the whole discourse. Changing of phrase translations, phrase order swapping and resegmentation are used to change the state and find the local optimum. Since this approach depends on the initial state and only finds local optima, it is somewhat unstable, but experiments show that the translation performance is comparable to that of beam search translation.

4.4 Evaluating Pronoun Translation

General purpose MT evaluation metrics such as BLEU measure the overall quality of translation output. When working on the translation of pronouns, only very few words are affected. BLEU, the de-facto standard evaluation metric, measures performance in terms of n-gram coverage. Since pronouns only make up a small percentage of words in the text and a wrong pronoun does not usually change the words surrounding the pronoun, BLEU will not reflect even large improvements

in pronoun translation quality and is therefore unsuitable for evaluating pronoun translation [LK10, HF10].

In order to measure their system's performance, Hardmeier et al. [HF10] therefore propose a precision/recall based measure: For each pronoun in the source text, they use word-alignments to retrieve its reference words R and translation path and phrase table information to retrieve the hypothesis words C. Inspired by BLEU they clip particular words in C at the value of their occurrence in R and then compute precision and recall in the following way:

$$Precision = \frac{\sum_{w \in C} c_{clip}(w)}{|C|} \qquad Recall = \frac{\sum_{w \in C} c_{clip}(w)}{|R|}$$

However, this metric has serious drawbacks [Har12]: It assumes that the pronoun in the hypothesis should be the same as the pronoun in the reference. But if the MT system chooses a different (correct) translation for the antecedent, then the correct pronoun might also differ from the reference. Guillou [Gui12] also mentions that this metric is ill-suited for highly inflective languages such as Czech.

A metric should therefore check if the target language pronoun agrees with its antecedent, for the pronoun needs to agree with its antecedent, even if the MT system chose an incorrect antecedent. This idea matches the linguistic requirements and should therefore be desired. But while this works well with hand-annotated anaphora resolution [Gui12], it seems to be difficult or even impossible with the currently available tools for automatic anaphora resolution [Har12]. Since automatic anaphora resolution has to be employed for all practical purposes, this evaluation idea cannot currently be used in practice on a large scale.

BLEU's unsuitability to measure changes to few words is also a problem in the field of discourse connectives [MPBHG12]. For this reason Meyer et al. [MPBHG12] propose a new family of metrics to measure performance of discourse connective translation. As the metric proposed by Hardmeier et al. [HF10] it compares reference and hypothesis: it employs a combination of word alignment and translation dictionary to spot the translation of source words, and then assigns each word to one of the classes *identical translation*, *equivalent translation*, and *incompatible translations*. Each member of the family of metrics then applies a slightly different formula on these values, including one that that is semi-automatic and includes human labelling of inserted connectives. While the authors receive good results for their context, the above criticism for the method by Hardmeier et al. [HF10] also applies here.

5. Resources

In this chapter we describe the data used in this work. In addition to the translation tasks that we work on, we describe the existing tools for part-of-speech tagging and anaphora resolution which we use. Furthermore we describe our own method of translating and evaluating anaphora resolution.

5.1 Translation Tasks

We work with two different English–German translation tasks covering two different domains:

news texts (news) are precise written texts for news presentation.
The translation system is a competitive system used in WMT [NZM+12]. It was trained on European Parliament Proceedings and the News Commentary data. The test set used is news2011.

TED talks (TED) are spontaneous speech from oral presentations.
The translation system is trained on European Parliament Proceedings, News Commentary data, the Common Crawl corpus and TED talks. Development set and est set are TED talks only.

For the analysis (Chapter 6) we need to create manual references. For this reason we limit our testset to 658 lines (of 3003) for news and 231 lines (of 1565) for TED.

5.2 Part-of-Speech Tags

For anaphora resolution evaluation (Chapter 5.4) and analysis (Chapter 6) part-of-speech (POS) tags for translation output are needed. We use two different methods to obtain these tags:

pos.text: The whole translation hypothesis text is processed by a part-of-speech tagger.

pos.pt: The phrase table contains part-of-speech tags for the words in a phrase. We extract these tags from the phrases from which the translation hypothesis was created.

The two methods do not provide the same tagging: The pos.text tags are created on the translation output. Since this text does not always consist of grammatically sound sentences, there will be errors in the tags. The pos.pt tags are created on the training sentences, i.e. grammatically sound sentences created by human writers. Yet since words may fulfil different roles in the translation text than they did in the training text, their POS tags may not match the translated text. The errors in the tags will not be the same for the two methods.

5.2.1 Part-of-Speech Taggers

For the different languages and different uses we used different part-of-speech taggers.

BLLIP parser is a two stage POS tagger and POS parse tree builder that can also provide information about syntactic and semantic heads. In a first stage the parser uses the Charniak parser's generative probabilistic model [Cha00] to produce the 50 best parses. In the second stage the parser finds the best parse in this list using a regularised maximum entropy reranker [CJ05]. The parser is also known as Charniak-Johnson parser or Brown reranking parser[1].

RFTagger is a tagger for fine-grained POS tags [SL08] that views POS tags as a sequence of attributes. Hidden Markov Models (HMMs) are used to split context probabilities into products of attribute probabilities. These probabilities are provided by decision trees[2].

TreeTagger is a Markov Model tagger for POS tags [Sch95] that uses decision trees for context restriction and a suffix lexicon to handle unknown words[3].

5.2.2 Finegrained POS Tags for German

We use the RFTagger [SL08] to create fine-grained POS tags for German texts and these tags are also used in the phrase table. For the words that are relevant to pronominal anaphora, the RFTagger output can take the forms shown in Table 5.1. The possible values for the placeholders ⟨·⟩ are given in Table 5.2. The symbol * is a wildcard.

regular noun	N.Reg.⟨case⟩.⟨number⟩.⟨gender⟩
names	N.Name.⟨case⟩.⟨number⟩.⟨gender⟩
personal pronouns	PRO.Pers.Subst.⟨person⟩.⟨case⟩.⟨number⟩.⟨gender⟩
possessive pronouns	PRO.Poss.Attr.-3.⟨case⟩.⟨number⟩.⟨gender⟩

Table 5.1: Forms of RFTags.

Personal pronouns only need to agree with the antecedent, so agreement can be established by simply checking whether both are plural, or if both are singular and their genders agree.

[1] The BLLIP parser is available online at `https://github.com/BLLIP/bllip-parser` with models for English and Chinese

[2] The RFTagger is available online at `http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/` with models for German, Czech, Slovene, Slovak and Hungarian

[3] The TreeTagger is available online at `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/` with models for a wide range of languages

⟨case⟩	Nom, Acc, Gen, Dat, *
⟨number⟩	Sg, Pl, *
⟨gender⟩	Masc, Fem, Neut, *
⟨person⟩	1, 2, 3, *

Table 5.2: Values for the placeholders in Table 5.1.

Possessive pronouns, on the other hand, need to agree with the antecedent (the "possessor") and the object that is being possessed. Unfortunately, the RFTagger's tags only give information about the possessed object and not about the antecedent. Therefore we applied the following rules to infer the antecedent's person, number and gender to create more finegrained tags for possessive pronouns:

word starting with	**person**	**number**	**gender**
mein	1	Sg	*
uns	1	Pl	*
dein	2	Sg	*
eu	2	Pl	*
sein	3	Sg	Masc / Neut
ihr	3	Sg / Pl	Fem / *

The symbol "*" stands for an any value, while "/" denotes the ambiguity that the value cannot be inferred from the word form but can be one of the list. So words starting with *sein* are tagged as .3.Sg.Masc/Neut., meaning they can either be .3.Sg.Masc. or .3.Sg.Neut.. Words starting with *ihr* are tagged as .3.Sg/Pl.Fem/*. meaning either .3.Sg.Fem. or .3.Pl.*..

The word *sein*, for example, may have the RFTag PRO.Poss.Attr.-3.Nom.Sg.Masc. Its more finegrained form is PRO.Poss.Attr.3.Nom.Sg.Masc/Neut.Sg.Masc .

5.3 Anaphora Resolution

Resolving anaphora is the process of identifying pronominal anaphors and their antecedents. We use the term *anaphora resolution pair* or just *pair* to describe the pair of an antecedent and the anaphor referring to it. When resolving pairs, our target is a list of these pairs in which each pair consists of one antecedent word and one anaphor word.

For reference we manually resolve anaphora. For each third person pronoun we find the antecedent and add the antecedent–anaphor pair to our list. If we cannot identify the antecedent (e.g. because the pronoun is used pleonastically), we do no create a pair for this pronoun.

For automatic anaphora resolution in English texts we conducted a number of preliminary experiments with a variety of tools. The JavaRAP[4] tool [QKC04] provided the best results for our tasks. This tool implements the rule-based *Resolution of Anaphora Procedure (RAP)* [LL94] for third person pronouns in English text.

[4]JavaRAP is available online at `http://wing.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html`

JavaRAP's version of the RAP algorithm (see Figure 5.1) creates a POS parse tree for the input text. In its vanilla version JavaRAP uses the Charniak parser [Cha00] to create these trees. We adapted it to use the two-stage BLLIP parser (see Chapter 5.2.1, [CJ05]) which achieves better parse results. With this adaptation JavaRAP also produces better anaphora resolution results. From the parse tree the algorithm extracts all the noun phrases, third person pronouns and reflexive pronouns together with a number of features such as agreement and occurrence information. A set of rules filters out pleonastic pronouns. For each of the remaining pronouns JavaRAP then creates a list of candidate antecedents from the noun phrases in the three sentences preceding the pronoun. These pairs of antecedent candidate and anaphor are then processed by one of two sets of grammatical rules: a *syntactic filter* is applied to third person pronouns and an *anaphor binding algorithm* is applied to lexical anaphors. These rule sets use some of the features extracted from the parse tree. For the remaining antecedent candidate anaphor pairs the algorithm uses the remaining features to calculate salience weights. Finally it chooses the best ranking antecedent candidate as the antecedent.

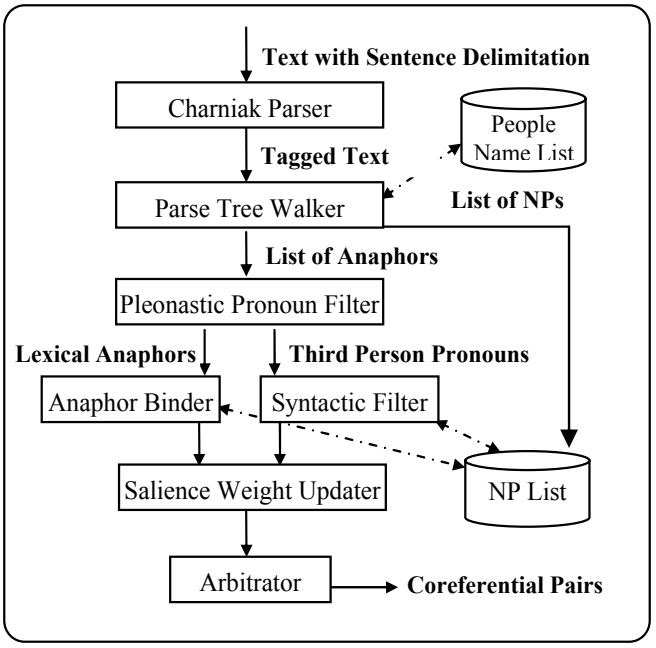


Figure 5.1: The JavaRAP procedure (from [QKC04]).

JavaRAP provides whole antecedent noun phrases. Yet for our task we only need the one antecedent word. Therefore, we use the vanilla version of the BLLIP parser to find the semantic head of the antecedent phrase. In the following steps we only use this semantic head and not the whole phrase.

Furthermore, we resolve anaphora chains: When the antecedent of one pair occurs as the anaphor in another pair, i.e. we have an anaphora chain, we follow the chain to its head. The head of such a chain is the antecedent that does not occur as an anaphor in any other pair. We then set the head of the chain as the antecedent for every pair in the chain.

Finally, we remove cyclic references, self references and entries for pleonastic pronouns.

5.4 Resolution Translation and Evaluation

Anaphora are translated correctly if the reference to their antecedent still holds true in the target language. Therefore we resolve anaphora resolution pairs in the source language and transfer them to the target language. There we check whether the anaphor still validly refers to the antecedent.

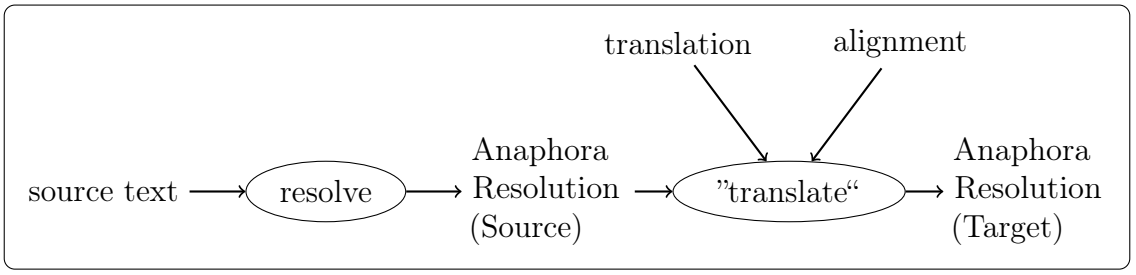
The procedure described in the following steps 1 – 3 provides a fully automatic mechanism that evaluates how well anaphora are translated. Step 1 together with the alternative steps 2 and 3 provide a semi-automatic scoring mechanism.

We use this procedure to translate both the manual and the automatic anaphora resolution lists.

Step 1 – Anaphora Resolution Translation

In order to translate the anaphora resolution pairs, we need to know how antecedent words and anaphor words were translated. This knowledge is represented in a word-level alignment between source text and translation output. For each sentence the phrase table entries for the phrase pairs used to create this translation are extracted from the phrase table. Each phrase table entry contains word level alignments between the words in the phrase pair, which we can now extract to create the alignment between source text and translated text.

Using this alignment, the (post-processed) anaphora resolution can now be translated by replacing each word with the word that it is aligned to.



Step 2 – Part-of-Speech Tagging

We are not particularly interested in the translated anaphora resolution. Our goal is to determine whether the anaphor still validly refers to the antecedent. We use part-of-speech (POS) tags for this test: the translated text is tagged with fine-grained POS tags (see Chapter 5.2). Then each word in the translated anaphora resolution is replaced by its finegrained POS tag .

Step 3 – Scoring

An anaphor validly refers to its antecedent if antecedent and anaphor agree in number and gender. We can now check if the POS tags for number and gender agree and calculate the percentage of anaphora resolution pairs for which there is agreement.

Alternative Step 2 & 3 – Manual Reference

Instead of checking agreement using POS tags, we can also manually correct the pronouns in the translated anaphora resolution and use this as a reference against which we compare the actual anaphora resolution translation.

5.5 Sources of Error

Throughout this work a number of tools are used that automatically process natural language. As these tools hardly ever produce perfect output, they introduce errors which accumulate as several tools are used together. All parts of this work are influenced by the errors introduces by at least some of the following sources of error:

anaphora resolution
: Errors in anaphora resolution can take several forms. The most common problem is that the anaphor does not actually refer to the antecedent. But it is also possible that words that can not be an antecedent/anaphor are chosen as antecedent/anaphor.

semantic head tagging
: The tagger does not always identify the correct semantic head, especially in nested clauses.

alignment
: The alignment for the words in phrases may not be accurate. Words can be misaligned or even incorrectly stay unaligned.

POS tagging
: Automatic POS taggers can provide erroneous tags. If the text to be translated is translation output that does not have perfectly sound grammar, the taggers will introduce additional errors. Furthermore, names are often tagged incorrectly.

6. Analysing Pronominal Anaphora

In the translation of a specific phenomenon like pronominal anaphora we can encounter two cases: either the implicit handling of this phenomenon by the current system is sufficient or we need to introduce explicit handling. Therefore, as a first step we analyse whether explicit handling of pronominal anaphora is necessary for English – German translation. We do this by studying the occurrence of pronouns in English and their translation into German. These studies include investigating the influence pronouns have on translation performance and what performance we could achieve if all pronouns were translated correctly.

6.1 Pronominal Anaphora in Text

In our testsets for news and TED we resolve anaphora both manually (.a.manual) and automatically with the JavaRAP tool (.a.auto) and perform post-processing as described in Chapter 5.3. We then look up the translations of the post-processed anaphora pairs as described in Step 1 in Chapter 5.4. We call the resulting target language anaphora resolution pairs *translated pairs*.

We define an antecedent–anaphor pair to be *correct* if the anaphor correctly refers to the antecedent, and *correctly translated* if the translated pair is correct.

In our analysis the manually created anaphora resolution .a.manual serves as a reference. The original .a.auto produced by the anaphora resolution tool may contain entries that are not anaphora pairs (e.g. a verb is incorrectly identified as antecedent or anaphor) and is therefore manually filtered to contain only correct pairs. This filtering that removes about a half of the anaphora pairs could also be done automatically by checking if anaphor and antecedent follow the required agreement requirements. For both .a.manual and .a.auto we look up the translations in our baseline translation system output. Then we apply two different filterings: First we filter out all pairs that are not translated into a pair (.a.manual.correctPair and .a.auto.correctPair). We do this because we only investigate pronouns that are translated into pronouns. In this study we are not interested in pronouns that are translated into grammatical structures that do not involve pronouns. Since errors may be introduced in the

translation, we filter further to keep only pairs that are correct in the target language (.a.manual.correctPair.correctTranslation and .a.auto.correctPair.correctTranslation).

Tables 6.1 and 6.2 show the numbers of anaphora resolution pairs found in these different anaphora resolution lists. For each automatic anaphora resolution list we include precision (P), recall (R) and f-score (F1) as compared to the corresponding manual anaphora resolution list.

anaphora list	**number of pairs**	**P**	**R**	**F1**
news.a.manual	288			
news.a.manual.correctPair	249			
news.a.manual.correctPair.correctTranslation	213			
news.a.auto	368	0.40	0.51	0.44
news.a.auto.correctPair	147	1.00	0.51	0.68
news.a.auto.correctPair.correctTranslation	114	0.78	0.54	0.63

Table 6.1: The size of anaphora resolution lists for news.

anaphora list	**number of pairs**	**P**	**R**	**F1**
ted.a.manual	170			
ted.a.manual.correctPair	161			
ted.a.manual.correctPair.correctTranslation	137			
ted.a.auto	176	0.47	0.48	0.47
ted.a.auto.correctPair	82	1.00	0.48	0.65
ted.a.auto.correctPair.correctTranslation	71	0.87	0.52	0.65

Table 6.2: The size of anaphora resolution lists for TED.

The post-processed and filtered automatic anaphora resolution achieves a similar performance on both our tasks. The automatic tool finds more pairs than the manual resolution, yet more than half of the pairs that the automatic tool finds are not correct pairs. Once we have filtered out all the pairs that are not correct, the resulting anaphora list has far fewer pairs than the manual reference. The recall for all lists is in the area of 0.5 which means that if we use the filtered automatic lists we can only work with about a half of the pairs that are actually in the text.

The precision for the .a.auto.correctPair lists is 1.0. This means that in those cases where the automatic anaphora resolution provides pairs of antecedent and anaphor that are translated into pairs, all of these pairs are really correct pairs that also occur in the manual list.

Whenever we use the .a.manual anaphora lists for experiments, these experiments are oracle experiments. The experiments with the a.auto anaphora lists are real experiments. In the following chapters we will not state this fact every time we conduct experiments.

6.2 Intra-Sentential and Inter-Sentential Anaphora

Anaphora can be divided into two groups: intra-sentential and inter-sentential (see Chapter 3.1). Tables 6.3 and 6.4 show how the anaphora resolution pairs from the above anaphora resolution are sorted into these groups and what percentage of these groups is translated correctly.

anaphora list	intra pronouns	inter pronouns	intra correct	inter correct
news.a.manual	50.7	49.3		
news.a.manual.correctPair	49.8	50.2		
news.a.manual.correctPair.correctTranslation	50.2	49.8	73.3	74.6
news.a.auto	63.3	36.7		
news.a.auto.correctPair	59.9	40.1		
news.a.auto.correctPair.correctTranslation	59.6	40.4	77.3	78.0

Table 6.3: Intra- and inter-sentential anaphora in news (in %).

In news, about half of the anaphora resolution pairs are intra-sentential and the other half is inter-sentential (news.a.manual). The automatic anaphora resolution (news.a.auto) finds a higher proportion of intra-sentential pairs. The translation performance is about the same for intra- and inter-sentential pairs.

anaphora list	intra pronouns	inter pronouns	intra correct	inter correct
ted.a.manual	24.1	75.9		
ted.a.manual.correctPair	24.2	75.8		
ted.a.manual.correctPair.correctTranslation	21.9	78.1	73.2	82.9
ted.a.auto	54.5	45.5		
ted.a.auto.correctPair	41.5	58.5		
ted.a.auto.correctPair.correctTranslation	40.8	59.2	87.5	85.3

Table 6.4: Intra- and inter-sentential anaphora in TED (in %).

In TED we find that only about a quarter of pairs is intra-sentential (ted.a.manual). This is because the sentences are shorter and the oral style of the text includes more inter-sentential reference than the written news. The automatic resolution again finds a higher proportion of intra-sentential pairs than we found manually (ted.a.auto). Translation performance on the manually identified pairs is better for inter-sentential pairs, while there is no performance difference for the automatically resolved anaphora resolution pairs.

Comparing the .correctPair lists, Tables 6.1 and 6.2 show that all of the automatically resolved pairs also occur in the reference. In terms of intra- and inter-sentential pairs we observe that the automatic tool misses more inter-sentential pairs than intra-sentential pairs.

6.3 Translation of Source Pronouns

We analyse how well individual source language pronouns are translated. We first analyse how often each source language pronoun occurs in our testsets and how many of these pronouns are already translated correctly. The results presented in Tables 6.5 and 6.6 help us determine how hard each individual pronoun is to translate.

source pronoun	occurrences	translated correctly
personal pronouns nominative		
he	49	100.0%
it	42	47.6%
she	10	90.0%
they	47	97.9%
personal pronouns objective		
her	6	100.0%
him	5	100.0%
them	11	100.0%
possessive pronouns		
his	21	100.0%
its	14	71.4%
their	44	88.6%

Table 6.5: Translations for news.

source pronoun	occurrences	translated correctly
personal pronouns nominative		
he	52	100.0%
it	36	47.2%
she	1	100.0%
they	28	100.0%
personal pronouns objective		
him	15	100.0%
them	3	100.0%
possessive pronouns		
his	14	100.0%
its	11	54.5%
their	1	100.0%

Table 6.6: Translations for TED.

For this analysis we use the a.manual.correctPair anaphora lists because we only investigate the pairs in which the source pronouns are actually translated into target

side pronouns. Since the numbers are from the manually created reference, they describe which target side pronoun the source side pronoun should be translated to.

We observe for both tasks that we already translate most pronouns correctly. However, the words *it* and *its* are only translated correctly in about half the cases. This suggests that the translation is harder for these pronouns than for the other pronouns.

In order to find the reason for this we analyse into which German pronouns the English pronouns are translated. If, for example, a source side pronoun is always translated to the same target side pronoun in our data, then we assume that the translation of this pronoun is easier than the translation of a pronoun that is translated to three different target side pronouns equally often.

The results of this second analysis are presented in Tables 6.7 and 6.8.

source pronoun	**target pronoun**	**how often**	**translated correctly**
personal pronouns nominative			
he	er	100.0%	100.0%
it	er	22.2%	12.5%
	es	44.4%	100.0%
	ihn	5.6%	0.0%
	sie	27.8%	0.0%
she	sie	100.0%	100.0%
they	sie	100.0%	100.0%
personal pronouns objective			
him	ihm	26.7%	100.0%
	ihn	73.3%	100.0%
them	ihnen	33.3%	100.0%
	sie	66.7%	100.0%
possessive pronouns			
his	sein	14.3%	100.0%
	seine	42.9%	100.0%
	seinem	14.3%	100.0%
	seinen	7.1%	100.0%
	seiner	21.4%	100.0%
its	ihr	9.1%	0.0%
	ihre	9.1%	100.0%
	ihren	63.6%	57.1%
	seine	9.1%	0.0%
	seinen	9.1%	100.0%
their	ihren	100.0%	100.0%

Table 6.7: Translations for TED.

source pronoun	target pronoun	how often	translated correctly
personal pronouns nominative			
he	er	100.0%	100.0%
it	er	19.0%	0.0%
	es	40.5%	76.5%
	ihn	7.1%	0.0%
	sie	33.3%	50.0%
she	er	10.0%	0.0%
	sie	90.0%	100.0%
they	es	2.1%	0.0%
	sie	97.9%	100.0%
personal pronouns objective			
her	ihr	16.7%	100.0%
	ihre	50.0%	100.0%
	ihrem	16.7%	100.0%
	ihren	16.7%	100.0%
him	ihm	60.0%	100.0%
	ihn	20.0%	100.0%
	seiner	20.0%	100.0%
them	sie	100.0%	100.0%
possessive pronouns			
his	sein	14.3%	100.0%
	seine	28.6%	100.0%
	seinem	9.5%	100.0%
	seinen	33.3%	100.0%
	seiner	14.3%	100.0%
its	ihrem	7.1%	100.0%
	ihren	7.1%	0.0%
	ihrer	7.1%	100.0%
	sein	14.3%	100.0%
	seine	28.6%	75.0%
	seinen	28.6%	75.0%
	seiner	7.1%	0.0%
their	ihr	6.8%	100.0%
	ihre	45.5%	100.0%
	ihrem	13.6%	100.0%
	ihren	4.5%	100.0%
	ihrer	15.9%	100.0%
	ihres	2.3%	100.0%
	seine	4.5%	0.0%
	seinem	4.5%	0.0%
	seinen	2.3%	0.0%

Table 6.8: Translations for news.

Tables 6.5 and 6.6 suggested that the word *it* is harder to translate since it is the only pronoun for which large percentages are translated incorrectly. In the results in Tables 6.7 and 6.8 we find that it has several possible translations unlike other pronouns such as *he*, *she* and *they* which almost always have the same translation. With over 40% of the cases the word *es* is the major target pronoun for *it*, and most of the translations into this target pronoun are correct. For TED, all occurrences of *it* that are to be translated to *es* are translated correctly. The remaining occurrences of the word *it* are mainly translated to *er* and *sie* with far lower accuracy.

For the word *it* the translation has a strong bias towards the word *es*. Even if it should be translated to something else, it is translated to *es* in most cases. Any approaches to handling anaphora explicitly should therefore improve the translation performance for the word *it* when translated to other words than *es* while keeping the translation performance of the other pronouns as good as it already is.

7. Post-Processing Based On Anaphora Resolution

The analysis has shown that not all pronouns are translated correctly. In this chapter we describe methods that process the translation system's output in order to correct those pronouns that were translated incorrectly.

Our post-processing methods are based on anaphora resolution lists. In Chapter 7.1 we describe how the anaphors are corrected in the anaphora lists of the translated text. We then present two post-processing methods in Chapters 7.2 and 7.3.

7.1 Correcting Translations of Anaphors

First, we create manual and automatic anaphora resolution lists on the source text and look up the translations as described in Chapter 5.4. We then use POS tags of antecedent and anaphora in the target language to produce anaphora resolution lists in which incorrectly translated pronouns are corrected.

For the correction we use the two different tagging methods pos.text and pos.pt described in Chapter 5.2 to provide POS tags for the translation output. We assume that both antecedent and anaphor POS tags are correct tags for the words they are tagging [1]. For each pair we check whether the agreement between anaphor and antecedent required by the English language is fulfilled. Wherever these requirements are not met, the anaphor does not refer to the antecedent which means it is an incorrect translation. In order to correct this pronoun, we change its POS tag so that it fulfils the agreement requirements with the antecedent's POS tag. Then we infer the correct pronoun word from the corrected pronoun tag.

In addition to these POS based methods, we also produce a manually corrected version of the anaphora resolution lists manual. This correction does not rely on POS tags. Instead the corrections are made directly from the word form of the text.

[1] This is certainly not the case for all words since we use automatic tagging methods which will almost always contain error. Therefore this correction also introduces new errors, but in the vast majority of the cases, the tags are correct and an actual correction takes place.

This set of corrections is a reference showing which words in a given list should be corrected. The results of the POS based correction methods can be compared to this to see what performance we can maximally achieve with the given anaphora resolution list. Figure 7.1 shows an example of the automatic correction method using POS tags and the manual method..

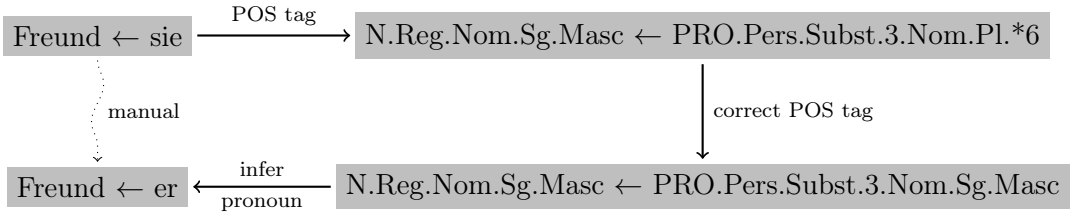


Figure 7.1: Example for the two correction methods.

7.2 Correcting Incorrect Pronouns

We use anaphora resolution to directly correct incorrect pronouns in the translation output: for each pair in the anaphora resolution list we compare the translated anaphor to the corrected anaphor translation. If the translated pronoun is incorrect, we replace this word by the correct pronoun word. In this way we change only pronoun words and leave the rest of the sentences as they were provided by the translation system. The overall change in there translation output therefore consists of only a few words.

As not all words are aligned and some anaphors are not aligned to pronouns, the numbers of correct translations in the previous chapter and the numbers of corrected words in this chapter do not add up to the total number of pronouns.

7.2.1 Changed Pronouns

First of all we look at the pronouns that are changed. For the manually created anaphora resolution list .a.manual.correctPair we used manual correction and the two POS based methods pos.text and pos.pt. For the automatically created list a.auto we cannot use manual correction because these lists contain entries in which the anaphor does not refer to the antecedent. Since these pairs are not correct, we cannot correct the words. Therefore we only use the automatic correction methods using POS tags for this anaphora resolution list. The results for the a.auto therefore show what changes we make if we perform the whole process automatically and do not filter out pairs that are not translated into pairs. Yet once we have filtered these lists into the lists that only contain correct pairs (.a.auto.correctPair) we can use the manual method as well as the two POS based methods. These results allow us to compare the quality of the changes provided by the POS based methods compared to the manual method.

We provide the numbers of pronouns that are changed, but since several pronouns can occur in the same sentence, we also look at how many sentences are affected by these changes. In addition investigate how many of the changed pronouns are intra- and inter-sentential. The results of this analysis are shown in Tables 7.1 and 7.2.

We notice that with the automatic lists .a.auto.correctPair less than half as many changes are made than with the manual lists .a.manual.correctPair. While there are

anaphora list	corrected by	words changed	lines changed	intra sentential	inter sentential
news.a.manual.correctPair	pos.text	64	53	48.4%	51.6%
	pos.pt	61	51	37.7%	62.3%
	manual	33	29	54.5%	45.5%
news.a.auto.correctPair	pos.text	31	27	51.6%	48.4%
	pos.pt	27	23	44.4%	55.6%
	manual	13	13	69.2%	30.8%
news.a.auto	pos.text	99	84	57.6%	42.4%
	pos.pt	88	73	53.4%	46.6%

Table 7.1: Changed pronouns in news.

anaphora list	corrected by	words changed	lines changed	intra sentential	inter sentential
ted.a.manual.correctPair	pos.text	32	30	18.8%	81.2%
	pos.pt	32	30	25.0%	75.0%
	manual	24	22	37.5%	62.5%
ted.a.auto.correctPair	pos.text	11	10	54.5%	45.5%
	pos.pt	10	9	50.0%	50.0%
	manual	8	7	50.0%	50.0%
ted.a.auto	pos.text	58	44	70.7%	29.3%
	pos.pt	58	43	70.7%	29.3%

Table 7.2: Changed pronouns in TED.

fewer entries in the automatic lists, this suggests that the pairs that the automatic system finds are more often translated correctly. For all experiments we observe cases in which several pronouns are corrected per sentence. This means that some sentences contain several incorrect pronouns. The results for intra- and inter-sentential anaphora show a mixed picture. In some experiments we have strong preference to one of the two types while for others the distribution of corrected pronouns is even. However, we perform post-processing in which we can access both antecedent and anaphor, even if they occur in different sentences, and our correction method is based only on the POS tags. Therefore, there is no difference in difficulty between intra- and inter-sentential anaphora with this correction approach.

In order to make the results more comparable, we evaluate which pronouns were corrected. We wish to find out which of the pronouns that are corrected really should be corrected, which of the pronouns that are corrected should not have been corrected, and which pronouns that should have been corrected are not corrected. This information is provided by precision (P), recall (R) and f-score (F1).

As references for calculating these scores we use two different lists of changed pronouns:

- The reference .a.manual is based on the manually created anaphora resolution list. Given this list we manually correct the incorrectly translated anaphora

resolution pairs. When comparing corrections to this reference, we compare two lists that have different entries. We observe which pronouns were corrected correctly and which necessary corrections were missed out.

- The reference manual is based on the anaphora resolution list that is currently being evaluated. Given this list we manually correct the incorrectly translated anaphora resolution pairs. The comparison with this reference shows which of the necessary changes that were possible with the given anaphora resolution list were actually made. When evaluating the manually created anaphora lists a.manual.correctPair this reference is the same as the reference .a.manual.

The results are presented in Table 7.3 for news and in Table 7.4 for TED.

anaphora list	corrected by	compared to	P	R	F1
news.a.manual.correctPair	pos.text	news.a.manual	0.47	0.91	0.62
		manual			
	pos.pt	news.a.manual	0.46	0.85	0.60
		manual			
news.a.auto.correctPair	pos.text	news.a.manual	0.39	0.47	0.43
		manual	0.32	0.92	0.47
	pos.pt	news.a.manual	0.41	0.44	0.42
		manual	0.32	0.85	0.47
	manual	news.a.manual	1.00	0.39	0.57
		manual	1.00	1.00	1.00
news.a.auto	pos.text	manual	0.19	0.58	0.29
	pos.pt	manual	0.19	0.52	0.28

Table 7.3: Precision, recall and f-score for changes on news.

anaphora list	corrected by	compared to	P	R	F1
ted.a.manual.correctPair	pos.text	news.a.manual	0.62	0.88	0.72
		manual			
	pos.pt	news.a.manual	0.66	0.88	0.75
		manual			
ted.a.auto.correctPair	pos.text	ted.a.manual	0.80	0.33	0.47
		manual	0.73	1.00	0.84
	pos.pt	ted.a.manual	0.73	0.33	0.46
		manual	0.80	1.00	0.89
	manual	ted.a.manual	1.00	0.33	0.50
		manual	1.00	1.00	1.00
ted.a.auto	pos.text	manual	0.17	0.42	0.24
	pos.pt	manual	0.17	0.42	0.24

Table 7.4: Precision, recall and f-score for changes on TED.

As we would expect, the manually created anaphora resolution list achieves better results than the automatically created lists. We also expected the reference manual

to be much closer to the actual changes than the a.manual reference which is shown by the higher f-score values for the manual reference.

The results show high recall values for reference manual which means that most of the corrections made are necessary corrections. For TED this recall is 1.0 which means that all changes were necessary changes. The resulting f-scores are very similar for the two references on news. This is due to the poor precision results. On TED, on the other hand, the comparison to manual achieves higher precision and thus much better f-scores.

The fully automatic unfiltered method achieves very low precision which means that only very few of the necessary changes were actually made. Recall has medium values, so about half of the changes made were necessary. This means that with the unfiltered anaphora resolution many changes are incorrect and only few of the necessary changes are made. However, since the filtering that created the a.auto.correctPair could also be performed automatically and we observe significantly better results for the filtered

		ted.a.manual.correctPair			ted.a.auto.correctPair			ted.a.auto	
original	corrected	pos.text	pos.pt	manual	pos.text	pos.pt	manual	pos.text	pos.pt
personal pronouns nominative									
er	es	2	2					11	11
	sie							5	5
es	er	5	5	7				6	6
	ihn	2	2	2	1	1	1	4	4
	sie	10	12	10	6	6	6	13	12
sie	er	1	1		1	1		1	1
	es	6	6						
personal pronouns other									
ihm	ihr							2	2
ihn	es	4			1			2	1
	sie							3	3
possessive pronouns									
ihre	ihre			1					
	seine	1	1	1	1	1	1	1	2
ihren	seinen				1	1		2	3
sein	ihr							1	1
seine	ihre							4	4
seinen	ihren	3	3	3				4	4

Table 7.5: Changed pronoun words in TED.

resolution lists, these results show that filtering the anaphora resolution lists is a necessary step for the automatic approach.

In addition the the numbers of changed pronouns and the comparison of actually corrected pronouns to those that should be corrected, we evaluate which pronoun words were changed into which other pronoun words. In Chapter 6.3 we analysed into which target language pronouns the given source language pronouns should be translated and into which they are translated by the baseline translation system. Now we analyse which words are recognised as translated incorrectly and which changes this results in.

		news.a.manual.correctPair			news.a.auto.correctPair			news.a.auto	
original	corrected	pos.text	pos.pt	manual	pos.text	pos.pt	manual	pos.text	pos.pt
personal pronouns nominative									
er	es	2	2	2	2	2	2	7	5
	sie	5	14		4	4		14	16
es	er	7	7	6	4	4	2	18	18
	ihn	1	1	2				1	
	sie	9	9	7	4	4	3	21	19
sie	er	13	6	3	1	1	1	9	6
	es	6	6	3	4	4	1	9	9
	sie			1				2	2
	ihn	2	1						
personal pronouns other									
ihm	ihnen							1	1
ihn	sie	1	1		1	1		1	1
ihnen	ihm							1	1
possessive pronouns									
ihre	seine	8	5	3	4	2		7	4
ihrem	seinem	2		1	1		1	1	
ihren	seinem			1			1		
	seinen	3	2	2	2	1	1	2	1
ihrer	seiner	1	1	1				1	1
seine	ihre	3	1		2	1		2	1
seinem	ihrem		1			1			1
seinen	ihren	1	3	1	2	2	1	2	2
seiner	ihrer		1						

Table 7.6: Changed pronoun words in news.

These words and their occurrences for each correcting method are shown in Table 7.6 for news and Table 7.5 for TED. The first line describes the anaphora resolution list used and the second line describes the correction method. The fifth column (news.a.manual.correctPair corrected by manual) shows which changes are actually necessary.

In Chapter 6.3 we saw that especially the word *it* is hard to translate correctly. It is often incorrectly translated as *es*. Here we see, that the corrections correspond to that observation and that most corrections are necessary for the word *es* and also that the most corrections are made for the word *es*.

7.2.2 BLEU Scores of Resulting Translation Text

Once we have corrected incorrect pronouns, we obtain a text in which pronouns are corrected, but the rest of the sentence remains as it was provided by the translation system. This text thus is a post-processed translation hypothesis and can be evaluated using MT evaluation metrics such as the de-facto standard BLEU. These scores tell us the upper bound for translation performance that can be achieved if all the changed pronouns are translated correctly according to the correction method used. We expect very small changes in BLEU score, as the mechanism of BLEU works in such a way that changes to a few individual words will not make a big difference for the score, even if they make a great difference for the meaning of the translation (Chapter 2.3).

		BLEU	
anaphora list	**corrected by**	**cs**	**ci**
news baseline	–	12.51	12.79
news.a.manual.correctPair	pos.text	12.44	12.72
	pos.pt	12.44	12.71
	manual	12.50	12.78
news.a.auto.correctPair	pos.text	12.47	12.75
	pos.pt	12.48	12.76
	manual	12.49	12.78
news.a.auto	pos.text	12.30	12.58
	pos.pt	12.33	12.60

Table 7.7: BLEU results for news.

The results for news in Table 7.7 show that the BLEU scores hardly change at all for the manually created anaphora resolution list and the automatically created list that was filtered for correct pairs. With the uncorrected list, on the other hand, we notice a performance decrease of 0.2 BLEU points. This means that the unfiltered lists contain so many incorrect antecedent-anaphor pairs that the correction introduces far more errors than corrections.

The results for TED in Table 7.8 show that we can achieve an improvement of about 0.2 BLEU points for the manually created lists, but hardly any improvements for the correct pairs in the automatic lists. For the whole automatic lists performance drops by just under 0.1 BLEU points.

anaphora list	corrected by	BLEU cs	BLEU ci
ted baseline	–	31.45	32.16
ted.a.manual.correctPair	pos.text	31.34	32.07
	pos.pt	31.55	32.27
	manual	31.62	32.35
ted.a.auto.correctPair	pos.text	31.51	32.24
	pos.pt	31.52	32.25
	manual	31.53	32.25
ted.a.auto	pos.text	30.40	31.13
	pos.pt	30.34	31.07

Table 7.8: BLEU results for TED.

7.2.3 Translation of Source Pronouns

We compare the pronoun translation accuracy of the translation system to the pronoun translation accuracy in the corrected translations. The pronoun translation of the translation system was analysed in detail in Chapter 6.3. As in that analysis we analyse how well the manually identified pronouns in .a.manual.correctPair are translated. For each pronoun in this anaphora list we provide the percentage that was translated correctly.

source pronoun	#	translated correctly	manual.correctPair correction pos.text	manual.correctPair correction pos.pt	auto.correctPair correction pos.text	auto.correctPair correction pos.pt
personal pronouns nominative						
he	49	100.0%	89.8%	71.4%	91.8%	91.8%
it	42	47.6%	92.9%	90.5%	73.8%	73.8%
she	10	90.0%	50.0%	40.0%	60.0%	60.0%
they	47	97.9%	80.0%	100.0%	97.9%	97.0%
personal pronouns objective						
her	6	100.0%	66.7%	66.7%	66.7%	66.7%
him	5	100.0%	100.0%	100.0%	80.0%	80.0%
them	11	100.0%	100.0%	100.0%	100.0%	100.0%
possessive pronouns						
his	21	100.0%	81.0%	71.4%	85.7%	85.7%
its	14	71.4%	100.0%	100.0%	85.7%	85.7%
their	44	88.6%	88.6%	95.5%	86.3%	88.6%

Table 7.9: Translations for news.

We observe high improvements for the word *it* as was already suggested by the number of changed pronouns in Tables 7.6 and 7.5. Unfortunately the performance of the other words drops slightly, so that the improvements achieved for the word *it* are accompanied by decreased performance for other pronouns. Therefore the overall translation accuracy does not improve as much as we might hope after the good improvement for the word *it*.

Comparing the changes using automatic and manual anaphora resolution input, we observe that the improvement for the word *it* is smaller for the automatic anaphora resolution list, but we do not see a drop in performance for the words *he* and *they*. This suggests that the filtered automatic anaphora lists do not contain all the occurrences of the anaphor *it* that need to be corrected.

source pronoun	#	**translated correctly**	**manual.correctPair correction** pos.text	pos.pt	**auto.correctPair correction** pos.text	pos.pt
personal pronouns nominative						
he	52	100.0%	96.2%	96.2%	100.0%	100.0%
it	36	47.2%	94.4%	88.8%	66.7%	66.7%
she	1	100.0%	0.0%	0.0%	0.0%	0.0%
they	28	100.0%	78.6%	78.6%	100.0%	100.0%
personal pronouns objective						
him	15	100.0%	73.3%	100.0%	93.3%	100.0%
them	3	100.0%	100.0%	100.0%	100.0%	100.0%
possessive pronouns						
his	14	100.0%	100.0%	100.0%	100.0%	100.0%
its	11	54.5%	100.0%	100.0%	54.5%	54.5%
their	1	100.0%	100.0%	100.0%	100.0%	100.0%

Table 7.10: Translations for TED.

7.3 N-Best Hypothesis Selection

The method we use to correct incorrectly translated pronouns (Chapter 7.1) has the disadvantage that incorrect POS tags can lead to erroneous changes to pronouns. When we directly change the words in the text according to these pronoun changes (Chapter 7.2), we also directly introduce these erroneous changes into the text. Therefore we develop a second post-processing method in which we use anaphora resolution to find a translation hypothesis in which the pronouns are correctly translated. In this method the changes that we can make to the pronouns are limited by the translation options of the decoder.

We start by creating anaphora resolution lists. We also obtain the n-best list with up to 300 hypotheses per sentence from the translation system. For each hypothesis we then check whether the anaphors in this hypothesis are translated correctly. Out of all the hypotheses in which the anaphor is translated correctly we select the hypothesis which has the highest translation probability (i.e. is the highest in the n-best list). If there is more than one anaphor in the sentence, we select the hypothesis in which the most anaphors are translated correctly. We use the hypothesis with the highest translation probability if more than one hypothesis has the same number of correct anaphors. If there is no hypothesis with correct pronouns we use the 1-best hypothesis.

7.3.1 Changed Pronouns

First of all we analyse which pronouns are changed by this method. We conduct these experiments only with .correctPair anaphora resolution lists, so we can use all three correction methods pos.text, pos.pt and manual. We report the number of sentences for that a different hypothesis is selected from the n-best list and how many pronouns this changes.

anaphora list	corrected by	words changed	lines changed
news.a.manual.correctPair	pos.text	23	21
	pos.pt	22	20
	manual	22	20
news.a.auto.correctPair	pos.text, pos.pt	14	12
	manual	3	3

Table 7.11: Results from selecting from n-best for news.

On the news task (Table 7.11) the result using pos.pt is exactly the same as using pos.text for the automatic anaphora resolution list. For the entries where the corrected anaphora resolution differs between pos.text and pos.pt, always one is correct in the first hypothesis and the other one is not in the hypotheses. Therefore they both result in the same texts.

As for the news task, the correcting methods do not all produce different outputs for TED (Table 7.12). On the contrary, for the automatic anaphora resolution the three methods produce exactly the same text, in which only two pronouns are changed.

anaphora list	corrected by	words changed	lines changed
ted.a.manual	pos.text	15	15
	pos.pt	17	17
	manual	14	14
ted.a.auto.correctPair	pos.text, pos.pt, manual	2	2

Table 7.12: Results from selecting from n-best for TED.

On both tasks, the automatic anaphora resolution list results in a lot less changes than the manual lists. We also observe that the number of pronouns that is corrected is significantly lower than the number of pronouns that should be corrected. We also note that far fewer pronouns are changed than with the other post-processing method (see Tables 7.3 and 7.4). This is due to the fact that if a pronoun should be corrected but the n-best list does not contain a sentence in which the corrected pronoun occurs, it cannot be corrected.

7.3.2 BLEU Scores of Resulting Translation Text

With this approach a different set of hypotheses form the translation than in the baseline translation system. In order to compare the overall translation quality of this method to the baseline we measure BLEU scores of the different sets of hypotheses. The hypotheses that we changed have a better pronoun translation, but the translation system originally ranked them as worse than its best hypothesis. The BLEU results will tell us, if and how much translation performance we lose with this approach.

We report the results of this post-processing method in Tables 7.13 and 7.14.

anaphora list	corrected by	BLEU cs	BLEU ci
news baseline	–	12.51	12.79
news.a.manual.correctPair	pos.text	12.49	12.78
	pos.pt	12.49	12.78
	manual	12.49	12.78
news.a.auto.correctPair	pos.text, pos.pt	12.49	12.77
	manual	12.49	12.77

Table 7.13: Results from picking from n-best for news.

On news we see similar results as for the pronoun correction approach: the BLEU score goes down slightly but only by negligible amounts.

On the TED task we also see results similar to the correcting pronouns approach. BLEU scores improve slightly, but not by more than 0.14 BLEU points which is not a significant improvement.

anaphora list	corrected by	BLEU cs	BLEU ci
ted baseline	–	31.45	32.16
ted.a.manual	pos.text	31.58	32.29
	pos.pt	31.59	32.30
	manual	31.57	32.28
ted.a.auto.correctPair	pos.text, pos.pt, manual	31.44	32.14

Table 7.14: Results from picking from n-best for TED.

On both tasks the result measured in BLEU is slightly better than for the correcting pronouns approach. For the news tasks this means the loss in performance is almost reduced to zero while for TED the average result improves slightly. However, the changes in BLEU compared to the baseline are so small that they are hardly noteworthy. This means that although we choose translation hypotheses that the translation system did not recognise as the best, we do not experience a loss in translation quality.

7.3.3 Translation of Source Pronouns

As for the pronoun correcting method, we analyse how well the pronouns are translated for comparison with the analysis in Chapter 6.3.

source pronoun	#	translated correctly	manual.correctPair correction pos.text	manual.correctPair correction pos.pt	auto.correctPair correction pos.text, pos.pt
personal pronouns nominative					
he	49	100.0%	100.0%	100.0%	100.0%
it	42	47.6%	81.0%	78.6%	69.0%
she	10	90.0%	90.0%	90.0%	90.0%
they	47	97.9%	97.9%	97.9%	97.9%
personal pronouns objective					
her	6	100.0%	100.0%	100.0%	100.0%
him	5	100.0%	100.0%	100.0%	100.0%
them	11	100.0%	100.0%	100.0%	100.0%
possessive pronouns					
his	21	100.0%	100.0%	100.0%	100.0%
its	14	71.4%	92.9%	92.9%	85.7%
their	44	88.6%	88.6%	88.6%	88.6%

Table 7.15: Translations for news.

We observe a good improvement for the translation of the word *it* and unlike the approach in the previous section (Tables 7.9 and 7.10), the performance for the other

source pronoun	#	translated correctly	manual.correctPair correction pos.text	manual.correctPair correction pos.pt	auto.correctPair correction pos.text, pos.pt
personal pronouns nominative					
he	52	100.0%	100.0%	100.0%	100.0%
it	36	47.2%	77.7%	72.2%	52.8%
she	1	100.0%	100.0%	100.0%	100.0%
they	28	100.0%	89.3%	89.3%	100.0%
personal pronouns objective					
him	15	100.0%	100.0%	100.0%	100.0%
them	3	100.0%	100.0%	100.0%	100.0%
possessive pronouns					
his	14	100.0%	100.0%	100.0%	100.0%
its	11	54.5%	72.7%	72.7%	54.5%
their	1	100.0%	100.0%	100.0%	100.0%

Table 7.16: Translations for TED.

pronouns does not go down. This means that in terms of overall pronoun translation performance the n-best list approach outperforms the pronoun correction approach. As the BLEU scores of the resulting texts are very similar, we can conclude that the approach using the n-best list is better than the pronoun correction approach.

Yet if we just look at the performance for the word *it*, we notice that the correcting approach provides better results. With the manual anaphora list that approach achieves a performance of up to 90% for the word it which is considerably better than for this approach. For the automatic anaphora resolution the difference between the two approaches is even bigger. While this approach achieves hardly any improvement for the word *it*, the correcting approach achieves some improvement. While the corrected anaphora lists contain a high number of corrected translations for the word *it*, this method is unable to find translation hypotheses with these corrected pronouns. This means that the translation system does not even have the correct pronouns among the 300 translation hypotheses used with the n-best selection method. Therefore it is unlikely to ever produce the correct translations and this approach will not be able to achieve huge improvements. Instead, the translation system's decoding needs to be influenced to produce the correct pronoun translations.

8. Discriminative Word Lexica for Pronouns

Phrase-based MT systems have rather limited context during decoding. As the antecedent of a pronoun is usually not found without these limits, we state in Chapter 3.4 that explicit anaphora handling is required. Following the fully explicit approaches in Chapter 7, we investigate how well explicit and implicit models using the Discriminative Word Lexicon (DWL, Chapter 2.2) handle anaphora. The DWL is a model that is not bound by the general limitations of a phrase-based system. It uses features from the whole input sentence and does not know about the segmentation of the sentence into phrases. Therefore it can leverage information from the whole sentence, which (depending on the task) often includes the antecedent of a pronoun.

When translating pronouns, the difference between a correct translation and an incorrect translation is usually just the one translated pronoun word. Therefore, if we can train DWL models such that out of all candidate pronouns the correct pronoun has the highest probability, we can boost the choice of the correct pronoun while still allowing the full flexibility of the decoder. In this chapter we investigate the use of extra features with the DWL. We aim to create a set of features that help the prediction of the probability of the translated pronouns. Some of the features use antecedents. As the post-processing in Chapter 7 they depend on anaphora resolution. In contrast, some features do not use this information but try to infer the correct pronoun choice implicitly. These features include different factors that correct pronoun translation depends on and thus might guide the model in selecting the correct word.

We use the MegaM toolkit [Dau04] to train the maximum entropy models for the DWLs[1]. This toolkit provides a fast training procedure using the Limited-Memory Broyden–Fletcher–Goldfarb–Shanno method (LM-BFGS).

We first describe base features and extra features in Chapter 8.1 and then report an evaluation of these features in Chapter 8.2.

[1] MegaM is available online at `http://www.umiacs.umd.edu/~hal/megam/`

8.1 Features for a Discriminative Word Lexicon

We investigate different features for the DWL models in order to find one with which the probabilities for correct pronouns is higher than for other pronouns. As base features we use *bag-of-words* [MHN09] and *bag-of-ngrams* [NW13] features and then add extra features.

bag-of-words

This feature type includes each word in the sentence once and thus enables the model to take into account information from the whole sentence and model long range dependencies. The order of the words is disregarded in this feature type.

bag-of-ngrams

While word order gets lost in the bag-of-words feature type, the bag-of-ngrams feature type takes into account the context in which the words occur in the source sentence. Instead of words, this feature type uses n-grams. Using $n = 1$ this feature type is the same as the bag-of-words. For $n > 1$ the n-grams include context around the words.

For this feature type we use all n-grams with $n \in 1, 2, 3$ of the source sentence.

8.1.1 Extra Features

The extra feature types are used in combination with the bag-of-words of bag-of-n-gram features. For each training example, the bag-of-words or bag-of-ngrams features are created and one or more extra feature types are added to it.

In Chapter 2.2 we describe how positive and negative training examples are created from the training data. We now describe the extra feature types and how we create the features for positive and negative training examples. In the evaluation, a feature is created in the same way as for positive training examples.

In some cases the same idea for a feature type results in two or three different features, since sometimes there is more than one way to create the feature for the training examples. Different versions of the same feature can be identified by the number appended to the feature type name. If only examples for which the extra feature can actually be created are used in training, we do not append a suffix to the name (e.g. PREVIOUS NOUNS). When we also use examples for which the extra feature cannot be created, we append a number 2 to the name of the feature (e.g. PREVIOUS NOUNS2). For features that use the translation of the antecedent, there also exists a third version. If we only use examples for which the antecedent is aligned (i.e. we can find the translation of the antecedent), we append a number 3 to the feature name (e.g. TARGET ANTECEDENT3).

Wherever an antecedent is required, we use automatic anaphora resolution with automatic post-processing on the training data to train the DWL model. This data contains quite a number of errors and is highly vulnerable since the antecedent is just one word. If the antecedent is identified incorrectly, the respective extra features will be trained with an incorrect antecedent.

ALIGN

This feature type includes the source word that the target word is aligned to in this sentence. With this extra feature the DWL knows which word is currently being translated.

positive example
The target word occurs in the reference, so we can look up and use the source word it is aligned to. We call this the source word that the target word is aligned to.

negative example
The target word occurs in the target vocabulary but not in the reference. This means there is a phrase pair for which the source phrase matches part of the source sentence and the target phrase contains the target word. So we can look up and use the source word that the target word is aligned to this phrase pair. We call this the source word that the target word would be aligned to.

If the target word is not aligned, we use an "unaligned" indicator as feature instead of the aligned word.

PREVIOUS NOUNS

This feature type indicates the nouns that come before the target word in the sentence. The translation of a pronoun depends on the translation of the antecedent. This word precedes the pronoun and is usually a noun. Therefore, it might be helpful to know which nouns there are in the source sentence for these could potentially be antecedents of the target word. However, since the antecedent does not have to be in the same sentence, this feature type does not necessarily include the antecedent. We say that a source word comes before the target word if its position is before the position the target word is aligned to. This position is determined as in the ALIGN feature type.

positive example
all nouns before the word that the target word is aligned to

negative example
all nouns before the word that the target word would be aligned to

PREVIOUS WORDS

This feature type is is a generalisation of the PREVIOUS NOUNS feature type. Instead of just nouns it indicates all the words that come before the current word in the sentence. As the antecedent may not be in the current sentence and therefore the PREVIOUS NOUNS feature type does not have the information it needs, the other words in the sentence may contribute helpful knowledge.

positive example
all words before the word that the target word is aligned to

negative example
all words before the word that the target word would be aligned to

TARGET PREVIOUS NOUNS

This feature type is similar to the PREVIOUS NOUN feature type, but instead of source nouns preceding the current word it includes the target nouns preceding the current word, i.e. the nouns that have already been translated. If the antecedent of the pronoun is in the same sentence, then it will be among these features.

positive example
all nouns before the target word

negative example
all nouns before the word that is aligned to the word the target word would be aligned to

TARGET PREVIOUS WORDS

This feature type is similar to the PREVIOUS WORDS feature type and a generalisation of the TARGET PREVIOUS NOUNS feature type. Instead of source words preceding the current word it includes the target words preceding the current word.

positive example
all words before the target word

negative example
all words before the word that is aligned to the word the target word would be aligned to

ANTECEDENT

The previous feature types try to address the problem of pronoun translation without explicitly handling pronouns. The ANTECEDENT feature type handles pronouns explicitly. If the word that the target word is aligned to or would be aligned to has an antecedent, we use the (source language) antecedent as a feature.

positive example
antecedent of the word that the target word is aligned to, if it has an antecedent

negative example
antecedent of the word that the target word would be aligned to, if it has an antecedent

TARGET ANTECEDENT

This feature type is similar to the ANTECEDENT feature type, but uses the target language antecedent instead of the source language antecedent. As the translation of a pronoun depends on the translation of the antecedent, this feature type should provide the model with all the information it needs to predict the correct pronoun translation.

positive example
if the word that the target word is aligned to has an antecedent, use the word aligned to this antecedent

negative example
if the word that the target word would be aligned to has an antecedent, use the word aligned to this antecedent (for training use the aligned reference word)

TARGET ANTECEDENT POS

This feature is similar to the TARGET ANTECEDENT feature, but uses the POS tag of the target antecedent instead of the target antecedent word. The idea behind this feature is that the translation of a pronoun depends on the part-of-speech of the antecedent rather than on the antecedent word itself. POS tags will also occur in the training data more often than the antecedent words themselves which may lead to more reliable models.

positive example
if the word that the target word is aligned to has an antecedent, use the POS tag of the word aligned to this antecedent

negative example
if the word that the target word would be aligned to has an antecedent, use the POS tag of the word aligned to this antecedent

TARGET ANTECEDENT GENDER

This feature type builds on the TARGET ANTECEDENT POS feature type. While the antecedent POS tags vary for different cases, the pronoun translation does not depend on the case. Therefore this feature type uses the gender of the antecedent if it is a noun and, as a fall-back, the antecedent POS tag if the antecedent is not a noun.

positive example
if the word that the target word is aligned to has an antecedent, use the gender of the word aligned to this antecedent as specified in the POS tag

negative example
if the word that the target word would be aligned to has an antecedent, use the gender of the word aligned to this antecedent as specified in the POS tag

8.2 Evaluation for Pronouns

In Chapter 6.3 we analyse into which target pronouns the source pronouns are translated and how often this translation is correct. Here we perform a similar analysis by evaluating the DWL models for each source pronoun individually. Chapter 6.3 shows that we only have room for improvement in the group of personal pronouns nominative with the word *it*. Therefore we limit the DWL evaluation to this group of pronouns (*he*, *she*, *it* and *they*) and their possible target pronouns *er*, *es*, *sie* and *ihn*.

As the basis of this evaluation we use the translation output hypothesis. Since this is the text in which we would like to improve the pronoun translation performance, we evaluate the anaphors in this text. For the feature types that include target language information, we use the part of the hypothesis before the current word which would also be available to the DWLs in decoding. As in Chapter 6.3 we use the .correctPair anaphora lists to select the sentences on which we evaluate the models. We also use these lists to provide antecedents. For source language POS tags we use the pos.text tags and for target language POS tags we use the pos.pt tags (Chapter 5.2).

For the evaluation we create the features for every sentence that contains an anaphor from the anaphora list. With these features we then evaluate the models for *er*, *es*, *sie*

bag-of-words + extra	he (49)	she (10)	it (42)	they (47)	all (148)
–	95.9	90.0	50.0	91.5	81.1
align	95.9	90.0	40.5	91.5	78.4
previous nouns	91.8	80.0	45.2	89.4	77.0
previous nouns2	93.9	80.0	42.9	91.5	77.7
previous words	89.8	80.0	40.5	87.2	74.3
previous words2	87.8	80.0	42.9	85.1	73.7
antecedent	89.8	80.0	33.3	89.4	73.0
antecedent2	93.9	80.0	42.9	91.5	77.7
align + previous nouns	95.9	70.0	42.9	91.5	77.7
align + previous words	89.8	80.0	40.5	78.7	71.6
align + antecedent2	98.0	90.0	38.1	91.5	78.4
align + previous nouns + antecedent2	98.0	80.0	42.9	91.5	79.0
align + previous words + antecedent2	91.8	90.0	40.5	93.6	77.7
target antecedent	87.8	80.0	38.1	93.6	75.0
target antecedent2	93.9	80.0	47.6	91.5	79.0
target antecedent3	81.6	80.0	40.5	89.4	72.3
target antecedent pos	87.8	80.0	50.0	91.5	77.7
target antecedent pos2	91.8	90.0	50.0	93.6	80.4
target antecedent pos3	87.8	80.0	50.0	89.4	77.0
target antecedent gender	85.7	80.0	57.1	85.1	77.0
target antecedent gender2	91.8	80.0	40.5	91.5	76.3
target antecedent gender3	85.7	80.0	50.0	80.8	73.7
target previous nouns	91.8	80.0	42.9	87.2	75.7
target previous words	81.6	80.0	40.5	83.0	70.3
antecedent2 + target antecedent2	91.8	80.0	47.6	91.5	78.4

Table 8.1: Results for news with base feature bag-of-words (in %).

and *ihn*. If the correct pronoun has the highest score, we mark it as correct, otherwise as incorrect. The results we present show percentages of how many pronouns were marked as correct.

The results for news are presented in Tables 8.1 and 8.2, the results for TED are shown in Tables 8.3 and 8.4. First, we look at the DWL without extra features (denoted by "–" in the tables). The DWL results are similar to the translation results in Chapter 6.3: high percentages for *he*, *she* and *they*, and a percentage close to 50% for *it*. The similarity between translation result and DWL result implies that simply training a model on the source words / n-grams (as it is done with DWLs) achieves the same amount of correctly translated pronouns as the full phrase based translation system. This also confirms that the word *it* is especially hard to translate and the features should especially aim at improving the translation of *it*. We also observe that for both data sets the models using bag-of-words features provide better results than the DWL using bag-of-ngrams features.

Comparing the two tasks, we notice a large difference between news and TED. For news (Tables 8.1 and 8.2) the extra features result in a loss of performance.

bag-of-ngrams + extra	he (49)	she (10)	it (42)	they (47)	all (148)
–	87.8	90.0	40.5	89.4	75.0
align	83.7	60.0	42.9	76.6	68.2
previous nouns	77.5	70.0	40.5	76.6	66.2
previous nouns2	91.8	90.0	23.8	89.4	71.6
previous words	85.7	80.0	33.3	83.0	69.6
previous words2	85.7	90.0	30.9	89.4	71.6
antecedent	79.6	60.0	30.9	80.8	64.9
antecedent2	85.7	80.0	38.1	83.0	71.0
align + previous nouns	83.7	80.0	40.5	78.7	69.6
align + previous words	85.7	90.0	33.3	83.0	70.3
align + antecedent2	83.7	60.0	42.9	80.8	69.6
align + previous nouns + antecedent2	81.6	60.0	40.5	76.6	66.9
align + previous words + antecedent2	83.7	80.0	28.6	80.8	66.9
target antecedent	77.5	70.0	33.3	83.0	66.2
target antecedent2	85.7	80.0	33.3	87.2	71.0
target antecedent3	71.4	70.0	35.7	85.1	65.5
target antecedent pos	75.5	70.0	40.5	89.4	69.6
target antecedent pos2	83.7	60.0	50.0	85.1	73.0
target antecedent pos3	75.5	70.0	40.5	89.4	69.6
target antecedent gender	69.4	70.0	35.7	87.2	65.5
target antecedent gender2	83.7	80.0	35.7	83.0	69.6
target antecedent gender3	73.5	70.0	33.3	91.5	67.6
target previous nouns	79.6	70.0	26.2	78.7	63.5
target previous words	77.5	80.0	33.3	74.5	64.2
antecedent2 + target antecedent2	85.7	70.0	38.1	83.0	70.3

Table 8.2: Results for news with base feature bag-of-ngrams (in %).

Whenever an additional feature type improves the performance for one pronoun, the performance for another pronoun drops at the same time. Consider for example the results for feature type bag-of-word + TARGET ANTECEDENT GENDER. The result for the word *it* improves a little, but the results for *he*, *she* and *they* deteriorate. Consequently the overall result is worse than the baseline.

The best results for news is with the TARGET ANTECEDENT POS2 feature type, but results are still worse than the baseline. In combination with the bag-of-words for 0.7% of the pronouns an incorrect target pronoun achieves the highest score. In combination with bag-of-ngrams the performance drops by 2.0%. Looking at the word *it* which is the one with the most room for improvement, we also do not observe any improvements. Instead performance drops dramatically by up to 16.7% for bag-of-words in combination with ANTECEDENT and bag-of-ngrams with PREVIOUS NOUNS2.

In contrast to the results on news, the extra features for TED either improve the result or leave it unchanged. In combination with the bag-of-words, the three features TARGET ANTECEDENT, TARGET ANTECEDENT3 and TARGET ANTECEDENT GENDER

bag-of-words + extra	**he (52)**	**she (1)**	**it (36)**	**they (28)**	**all (117)**
–	96.2	100.0	47.2	89.3	79.5
align	100.0	100.0	38.9	96.4	80.3
previous nouns	96.2	100.0	47.2	92.9	80.3
previous nouns2	94.2	100.0	47.2	89.3	78.6
previous words	100.0	100.0	44.4	89.3	80.3
previous words2	100.0	100.0	44.4	85.7	79.5
antecedent	100.0	100.0	44.4	92.9	81.2
antecedent2	96.2	100.0	47.2	89.3	79.5
align + previous nouns	88.5	100.0	44.4	100.0	77.8
align + previous words	98.1	100.0	47.2	89.3	80.3
align + antecedent2	100.0	100.0	47.2	89.3	81.2
align + previous nouns + antecedent2	100.0	100.0	47.2	92.9	82.0
align + previous words + antecedent2	98.1	100.0	47.2	89.3	80.3
target antecedent	100.0	100.0	47.2	92.9	82.0
target antecedent2	96.2	100.0	50.0	89.3	80.3
target antecedent3	100.0	100.0	47.2	92.9	82.0
target antecedent pos	100.0	100.0	44.4	92.9	81.2
target antecedent pos2	98.1	100.0	47.2	85.7	79.5
target antecedent pos3	100.0	100.0	44.4	92.9	81.2
target antecedent gender	100.0	100.0	47.2	92.9	82.0
target antecedent gender2	96.2	100.0	47.2	89.3	79.5
target antecedent gender3	100.0	100.0	44.4	89.3	80.3
target previous nouns	98.1	100.0	47.2	92.9	81.2
target previous words	90.4	100.0	41.7	92.9	76.1
antecedent2 + target antecedent2	96.2	100.0	50.0	89.3	80.3

Table 8.3: Results for TED with base feature bag-of-words (in %).

achieve a result of 82.0% which is a 2.5% improvement over the baseline. With the bag-of-ngrams, the feature type TARGET ANTECEDENT POS also achieves a result of 82.0%, but since the bag-of-ngrams baseline is worse than the bag-of-words baseline, this is an improvement of 7.6%.

While these are encouraging results, the source of the improvements is discouraging. In combination with bag-of-words the extra features with the best result improve only on *he* and *they* while *she* cannot be further improved. The performance for *it* remains unchanged. This means that the DWLs got even better at predicting the words they were already good at, but they did not improve for the word *it* for which the most improvement is needed and for which there is the most potential for improvement. Indeed, while the overall performance is never worse than the baseline, the performance for *it* often drops below baseline performance. The best result for *it* is in combination with the features TARGET ANTECEDENT2 and ANTECEDENT2 + TARGET ANTECEDENT2 which both achieve 50.0% for *it*, a 2.8% improvement. This is an improvement for this particular word, but since the results for the other words

bag-of-ngrams + extra	**he (52)**	**she (1)**	**it (36)**	**they (28)**	**all (117)**
–	88.5	100.0	50.0	78.6	74.4
align	96.2	100.0	50.0	85.7	79.5
previous nouns	88.5	100.0	47.2	82.1	74.4
previous nouns2	90.4	100.0	50.0	82.1	76.1
previous words	96.2	100.0	47.2	82.1	77.8
previous words2	96.2	100.0	50.0	85.7	79.5
antecedent	92.3	100.0	52.8	85.7	78.6
antecedent2	90.4	100.0	50.0	85.7	76.9
align + previous nouns	73.1	100.0	44.4	100.0	70.9
align + previous words	94.2	100.0	50.0	89.3	79.5
align + antecedent2	96.2	100.0	50.0	89.3	80.3
align + previous nouns + antecedent2	96.2	100.0	50.0	85.7	79.5
align + previous words + antecedent2	96.2	100.0	50.0	92.9	81.2
target antecedent	92.3	100.0	52.8	85.7	78.6
target antecedent2	92.3	100.0	50.0	82.1	76.9
target antecedent3	92.3	100.0	52.8	82.1	77.8
target antecedent pos	98.1	100.0	52.8	89.3	82.0
target antecedent pos2	94.2	100.0	50.0	89.3	79.5
target antecedent pos3	98.1	100.0	52.8	85.7	81.2
target antecedent gender	96.2	100.0	52.8	85.7	80.3
target antecedent gender2	90.4	100.0	50.0	82.1	76.1
target antecedent gender3	94.2	100.0	52.8	85.7	79.5
target previous nouns	88.5	100.0	50.0	85.7	76.1
target previous words	88.5	100.0	47.2	82.1	74.4
antecedent2 + target antecedent2	94.2	100.0	50.0	82.1	77.8

Table 8.4: Results for TED with base feature bag-of-ngrams (in %).

remain unchanged, their overall result is 80.3%, which is only a 0.8% improvement over the baseline.

In combination with bag-of-ngrams, the best extra feature type for overall performance is also among the best for the word *it*. Still, the performance for it only increases by 2.8% while the overall improvement is 7.6%. But in contrast to the bag-of-words base feature type, the other words also have more room for improvement with the bag-of-ngrams feature type.

9. Source Discriminative Word Lexica for Pronouns

The Discriminative Word Lexicon (DWL) is a binary classifier that predicts whether or not a target word is to be used in the translation (Chapter 2.2). For this prediction it uses one maximum entropy model for each target language word.

Just like the DWL the Source Discriminative Word Lexicon (SDWL) uses maximum entropy models to predict whether or not a target word is to be used in the translation. Yet while the DWL has one model for each target word, the SDWL has one model for each source word. These models are multiclass classifiers that assign a set of features to a class. We choose classes in such a way that each class corresponds to a target word: In Chapter 6.3 we observed that we only have room for improvement in the group of personal pronouns nominative with the word *it*. Therefore we train SDWLs for this group of pronouns (*he*, *she*, *it* and *they*) and create classes for their possible target pronouns *er*, *es*, *sie* and *ihn*.

As DWLs, the SDWLs use information from the whole sentence. An SDWL directly predicts the target word to be used as translation of a source word, so we do not need to evaluate many models and chose the one with the highest score. As the SDWL can also be included as one model in the log-linear model, so that score from the SDWL is combined with the full flexibility of the decoder.

9.1 Model Types

Since the SDWL is a multiclass model, we do not create positive and negative examples as for the binary DWL model, but assign each example to a class. Each time one of the investigated source language pronouns occurs in the training text, we look at the reference word that is aligned to the pronoun. Ideally we would like to create the features for the sentence and then assign the class that represents the aligned word. But since pronouns are not always translated into pronouns, there will be pronouns that are not aligned to a word represented by a class in the model. To reflect this, we create two different model types:

SDWL-4c

If the pronoun is aligned to one of the words represented by a class, we create the features and assign them to that class. If the pronoun is not aligned to a word represented by a class or not aligned at all, we do not create a training example from this pronoun occurrence.

This model has four classes: *er*, *sie*, *es* and *ihn*.

SDWL-5c

If the pronoun is aligned to one of the words represented by a class, we create the features and assign them to that class. Yet if the pronoun is aligned to a different word, we assign the features of the sentence to an additional *other* class. If the pronoun is not aligned, we do not create a training example from this pronoun occurrence.

This model has five classes: *er*, *sie*, *es*, *ihn* and *other*.

9.2 Features

As with the DWLs we can use different feature types to train and evaluate the SDWL models. We use the same *bag-of-words* [MHN09] and *bag-of-ngrams* [NW13] base features as for the DWLs.

bag-of-words

This feature type includes each word from the sentence once and thus enables the model to take into account information from the whole sentence and model long range dependencies. The order of the words does not play a role in this feature type.

bag-of-ngrams

While word order gets lost in the bag-of-words feature type, this feature type takes into account the context in which the words occur in the source sentence. Instead of a bag of words, a bag of n-grams is used. Using $n = 1$ this feature type is the same as the bag-of-words, but for $n > 1$ this feature includes context. We use the bag-of-ngrams feature with all $n \in 1, 2, 3$.

extra feature: target antecedent

This extra feature type includes the target language antecedent word. The idea behind this feature type is that the translation of a pronoun depends on the translation of the antecedent word. Therefore the knowledge of the translation of the antecedent should help the DWL predict the correct translation of the currently translated word.

If the word has an antecedent, we use the target word aligned to this antecedent as extra feature.

9.3 Evaluation for Pronouns

The evaluation of the SDWLs is very similar to the evaluation of the DWLs (Chapter 8.2). We evaluate the models for the same pronouns, but instead of evaluating all models for each word, we only evaluate the model for the current pronoun.

The results in Table 9.1 for news and Table 9.2 for TED include the results for both model types with the base features and in combination with the extra feature.

	he	**she**	**it**	**they**	**all**
bag-of-words	**(49)**	**(10)**	**(42)**	**(47)**	**(148)**
SDWL-5c	95.9	90.0	5.1	93.6	69.0
SDWL-5c + target antecedent	87.8	90.0	17.9	97.9	71.3
SDWL-4c	100.0	90.0	38.5	97.9	81.2
SDWL-4c + target antecedent	100.0	90.0	30.8	97.9	79.0
bag-of-ngrams					
SDWL-5c	95.9	90.0	10.3	97.9	71.8
SDWL-5c + target antecedent	95.9	90.0	12.8	97.9	72.6
SDWL-4c	100.0	90.0	38.5	97.9	81.2
SDWL-4c + target antecedent	100.0	90.0	43.6	97.9	82.7

Table 9.1: Results for news (in %).

In the results for the news task we observe that the models for *he*, *she* and *they* are very good. This was to be expected since Table 6.5 shows that most occurrences of these words are translated to the same words. The model for the word *it*, on the other hand, shows terrible results. In the SDWL-5c there are five classes: *er*, *sie*, *es*, *ihn* and *other*. So if the SDWL-5c would randomly chose a class, it had a 20% chance of choosing the correct pronoun. The results that we observe, however, are below this value of chance, and also well below the result achieved by the translation system. This is because in the SDWL-5c models the *other* class achieves the highest score in most of the cases. Using the SDWL-4c, which does not have the *other* class, improves the results considerably.

In the direction the SDWL is working, the context provided by n-grams seems to be useful information, since the models with the bag-of-ngrams perform slightly better than the models with the bag-of-words. The extra feature type TARGET ANTECEDENT also yields improvements. In most combinations it improves the performance of the model *it*, and leaves the performances of the other models unchanged. However, in combination with the SDWL-4c model and the bag-of-words base features the TARGET ANTECEDENT feature type leads to worse performance.

The best result is the SDWL-4c in combination with the bag-of-ngrams base feature and the TARGET ANTECEDENT extra feature with a result of 43.5% for *it* and 82.7% overall.

In the results for TED we also observe nearly perfect performance for the models *he*, *she* and *they*, which again was expected. But unlike on the news task, the performance

bag-of-words	**he (52)**	**she (1)**	**it (36)**	**they (28)**	**all (117)**
SDWL-5c	100.0	100.0	47.1	100.0	83.7
SDWL-5c + target antecedent	94.2	100.0	47.1	89.3	79.8
SDWL-4c	100.0	100.0	47.1	89.3	81.2
SDWL-4c + target antecedent	100.0	100.0	47.1	100.0	83.7
bag-of-ngrams					
SDWL-5c	100.0	100.0	44.1	96.4	81.2
SDWL-5c + target antecedent	100.0	100.0	44.1	100.0	82.8
SDWL-4c	100.0	100.0	47.1	100.0	83.7
SDWL-4c + target antecedent	100.0	100.0	47.1	92.9	82.0

Table 9.2: Results for TED (in %).

of the model for *it* is the same as the result from the translation system in Table 6.6. On this task, the difference between the SDWL-5c and the SDWL-4c model types and the bag-of-words and bag-of-ngrams base features are a lot smaller than for news. The TARGET ANTECEDENT feature type does not achieve big improvements as for news. But similar to news, the same combinations of model type and base feature result in a loss of performance in combination with the TARGET ANTECEDENT feature type. On TED this happens with the SDWL-5c model type in combination with the bag-of-words and the SDWL-4c model type in combination with the bag-of-ngrams.

10. Comparison of the Approaches

We investigate four approaches for explicit anaphora handling in machine translation: post-processing by correcting words (Chapter 7.2), post-processing by searching the n-best list (Chapter 7.3), translation prediction with the Discriminative Word Lexicon (Chapter 8) and translation prediction with the Source Discriminative Word Lexicon (Chapter 9).

In this chapter we compare the results of these approaches. In Chapter 6.3 we analyse into which target pronouns the source pronouns are translated and how often this translation is correct in the baseline translation. The results of that analysis show that the most improvement is necessary in the group of personal pronouns nominative with the word *it*. In order to set the improvements for this word into context we limit our evaluation to this group of pronouns: *he*, *she*, *it* and *they*. For each of these pronouns we calculate which percentage of these pronouns is correctly translated or, in the case of DWL and SDWL, correctly predicted. We report results for the pronouns in the anaphora resolution lists a.manual.correctPair. These are the pronouns for which we know that they are translated into pronouns and for which we have a manually created reference.

For this comparison we repeat the best results for each approach from the individual chapters. The anaphora resolution list on which we evaluate the results does not have to be the same as the one that provided the anaphora resolution pairs. While we always evaluate on a.manual.correctPair, we conducted experiments with both a.manual.correctPair and a.auto.correctPair as input. The results therefore state for each result the anaphora resolution list that was used to obtain it. In comparing the results from using a.manual.correctPair to the ones using a.auto.correctPair we can determine the performance difference between using manual and automatic anaphora resolution.

The results for news in Table 10.1 and for TED in Table 10.2 first of all show that despite the differences in the tasks, the baseline performance is essentially the same.

The results for the two tasks also have in common that the DWL and SDWL approaches are unable to produce results better than the translation baseline. These results do not show the percentages of correct translation using these models, they

	he (49)	**she** (10)	**it** (42)	**they** (47)	**all** (148)
baseline translation	100.0	90.0	47.6	97.9	83.8
post-processing – correcting words					
corrected by pos.text (.a.manual.correctPair)	89.8	50.0	92.9	80.0	84.9
corrected by pos.pt (.a.manual.correctPair)	71.4	40.0	90.5	100.0	83.8
corrected by pos.text (.a.auto.correctPair)	91.8	60.0	73.8	97.9	86.5
corrected by pos.pt (.a.auto.correctPair)	91.8	60.0	73.8	97.9	86.5
post-processing – n-best					
corrected by pos.text (.a.manual.correctPair)	100.0	90.0	81.0	97.9	93.2
corrected by pos.pt (.a.manual.correctPair)	100.0	90.0	78.6	97.9	92.6
corrected (.a.auto.correctPair)	100.0	90.0	69.0	97.9	89.9
dwl words					
baseline	95.9	90.0	50.0	91.5	81.1
target antecedent pos2 (.a.manual.correctPair)	91.8	90.0	50.0	93.6	80.4
target antecedent pos2 (.a.auto.correctPair)	91.8	80.0	42.9	93.6	77.7
dwl ngrams					
baseline	87.8	90.0	40.5	89.4	75.0
previous nouns2 (.a.manual.correctPair)	91.8	90.0	23.8	89.4	71.6
previous nouns2 (.a.auto.correctPair)	91.8	90.0	23.8	89.4	71.6
sdwl words (SDWL-4c)					
baseline	100.0	90.0	38.5	97.9	81.2
target antecedent (.a.manual.correctPair)	100.0	90.0	30.8	97.9	79.0
target antecedent (.a.auto.correctPair)	100.0	100.0	31.0	100.0	80.4
sdwl ngrams (SDWL-4c)					
baseline	100.0	90.0	38.5	97.9	81.2
target antecedent (.a.manual.correctPair)	100.0	90.0	43.6	97.9	82.7
target antecedent (.a.auto.correctPair)	100.0	100.0	37.9	100.0	82.4

Table 10.1: Pronoun evaluation results for news (in %).

show the percentages of correct prediction by these models. Since this prediction is not better than the results the translation system already achieves, we did not build these approaches into the decoder and therefore we do not have translations using these models. Comparing DWL and SDWL we see that the SDWL performs better than the DWL. This suggests that for the pronouns it is better to directly predict

	he (52)	**she (1)**	**it (36)**	**they (28)**	**all (117)**
baseline translation	100.0	100.0	47.1	100.0	83.7
post-processing – correcting words					
corrected by pos.text (.a.manual.correctPair)	96.2	0.0	94.4	78.6	90.6
corrected by pos.pt (.a.manual.correctPair)	96.2	0.0	88.8	78.6	88.9
corrected by pos.text (.a.auto.correctPair)	100.0	0.0	66.7	100.0	88.9
corrected by pos.pt (.a.auto.correctPair)	100.0	0.0	66.7	100.0	88.9
post-processing – n-best					
corrected by pos.text (.a.manual.correctPair)	100.0	100.0	77.7	89.3	90.6
corrected by pos.pt (.a.manual.correctPair)	100.0	100.0	72.2	89.3	88.9
corrected (.a.auto.correctPair)	100.0	100.0	52.8	100.0	85.5
dwl words					
baseline	96.2	100.0	47.2	89.3	79.5
target antecedent (.a.manual.correctPair)	100.0	100.0	47.2	92.9	82.0
target antecedent (.a.auto.correctPair)	100.0	100.0	44.4	92.9	81.2
dwl ngrams					
baseline	88.5	100.0	50.0	78.6	74.4
target antecedent pos (.a.manual.correctPair)	98.1	100.0	52.8	89.3	82.0
target antecedent pos (.a.auto.correctPair)	94.2	100.0	50.0	89.3	79.5
sdwl words (SDWL-4c)					
baseline	100.0	100.0	47.1	89.3	81.2
target antecedent (.a.manual.correctPair)	100.0	100.0	47.1	100.0	83.7
target antecedent (.a.auto.correctPair)	100.0	100.0	33.3	100.0	79.5
sdwl ngrams (SDWL-4c)					
baseline	100.0	100.0	47.1	97.9	83.7
target antecedent (.a.manual.correctPair)	100.0	100.0	47.1	92.9	82.0
target antecedent (.a.auto.correctPair)	100.0	100.0	33.3	100.0	79.5

Table 10.2: Pronoun evaluation results for TED (in %).

them from the source text rather than hoping for the right model to produce the highest score.

For the post-processing approaches we achieve improvements over the translation baseline. Since the results vary between the two tasks, we examine the results for each task individually.

On the news task we achieve small improvements using the correcting words approach. Interestingly, the results are better when we use the automatic anaphora resolution lists rather than the manually created anaphora resolution lists. This will be due to the fact that we use automatic POS tagging methods for the correction of the pronouns and that these tags contain errors. The pronouns in the automatic anaphora resolution list are identified correctly, so they tend to be easier cases for which there exist correct POS tags (recall that the RAP procedure described in Chapter 5.3 is based on POS tags). The manual anaphora resolution lists, on the other hand, contain pronouns that the automatic resolution did not correctly resolve. It is possible that the POS tags for these pronouns are incorrect and that through this new errors are introduced.

The n-best approach, on the other hand, achieves good improvements. This approach can only change a pronoun if there is a hypothesis with the changed pronoun in the n-best list. Thus it is limited to the translation hypotheses that the decoder created and cannot freely introduce new errors. With this approach we achieve a result of 93.2% correctly translated pronouns which is a 9.4% improvement overall and includes a 33.4% improvement for the word *it*.

On the TED task, the post-processing by correcting words and the post-processing on the n-best list produce very similar results. With the correcting words approach and manual anaphora resolution, the performance for the word *it* improves greatly while the performance for the other words drop slightly[1]. With the n-best approach, the performance for the word *it* does not improve as much, but the other words to not lose performance, so the overall result achieve similar results. The best result is 90.6% which is achieved by both post-processing methods using the manual anaphora resolution list and the pos.text POS tags.

The results that we compare here show that with an explicit handling of anaphora in MT we can significantly improve the number of pronouns that are translated correctly. The DWL and SDWL approaches, however, do not produce promising results, even if they use explicit anaphora resolution data.

[1]Please note that on the TED task there is only one occurrence of the word *she*, so a drop in performance from 100.0% to 0.0% means that only one pronoun has been translated incorrectly.

11. Conclusion

In this thesis we studied pronominal anaphora in English – German machine translation. We analysed the occurrence and translation of pronouns for news texts and TED talks. Conducting oracle experiments we found that our translation system handles pronouns insufficiently since about a quarter of the pronouns that take part in pronominal anaphora are translated incorrectly (Chapters 6.2 and 7.2.1). From this we conclude that the translation system is indeed restricted by the limitations of the independence assumptions inherent in phrase-based MT. Explicit handling of anaphora in machine translation is necessary.

From our experiments with automatic anaphora resolution we conclude that the quality of the results produced by the JavaRAP tool is insufficient. Yet if we build rule-based automatic filters that remove all the problematic pairs, the resulting anaphora pairs achieve good results. By filtering we will obtain a smaller number of anaphora pairs than the tool found, but for the filtered pronouns we obtain good results. This shows that we can use automatically identified anaphora pairs, if we ensure that only correct anaphora pairs remain in the list (Chapter 7.2).

We developed four approaches that handle anaphora explicitly. The first two are post-processing approaches that leverage lists of anaphora pairs. These approaches do not change the translation process itself but work with the translation results produced by the translation system. In the first approach we identify the pronoun words that were translated incorrectly and correct them in the text with a set of grammatical rules. With this approach we are able to improve the pronoun translation performance for the nominative pronouns by up to 6.3% for the oracle experiment and up to 5.6% for the real experiment depending on the task. In the second post-processing based approach we work with the translation system's n-best list. From this list we select the hypotheses in which most of the pronouns are translated correctly. Compared to the correcting method this approach gives better pronoun translation results on the news task and comparable results on the TED task. Both post-processing approaches improve the pronoun translation performance. In terms of overall translation performance measured in BLEU (Chapters 7.2.2 and 7.3), the result deteriorates by up to 0.08 BLEU points on the news task, which is

hardly a change at all. On the TED task the result improves by a very small amount of 0.2 BLEU points.

The other two approaches are build to work in the translation process and support the decoder while still allowing its full flexibility. The DWL uses features to predict for each target word whether or not it should be included in the translation. The SDWL uses similar features to directly predict the translation of a source word. For these approaches we developed a range of feature: some make use of anaphora lists while others do not. In the experiments the features which use anaphora list achieve the best results. This shows that the anaphora lists contain information that the models cannot implicitly infer from the other information. The results summary shows that even with this additional data these models are unable to predict pronouns better than the baseline translation system already does. Therefore, we did not build the DWL and SDWL approaches into the decoder but only evaluated the model's ability to predict the correct pronouns.

The results of the work presented in this thesis show that explicit anaphora handling is necessary for phrase-based machine translation. We developed two post-processing methods which are able to improve the pronoun translation performance and, in the case of the TED task, even improve the translation performance measured in BLEU. The two other approaches that predict pronouns given a set of input features are unable to predict the correct pronoun better than the baseline translation.

11.1 Outlook

In this thesis we worked with the English – German language pair. The approaches developed in this work should also work for other language pairs. It would be interesting to see how they perform on other language pairs and in how far the difficulties lie elsewhere.

We developed DWL and SDWL models and evaluated their power to predict pronouns. Their predictions are no improvement over the translations from translation system. However, in combination with the full flexibility of the decoder the predictive power of the DWL and SWDL may still yield improvements for the translation.

The n-best list based approach yields the most promising results. By choosing hypotheses in which more pronouns are correct from the n-best list, we actually improved the BLEU score on the TED task. Future research could investigate whether an n-best re-ranking based on pronominal anaphora and other discourse connectives could improve translation performance.

Nomenclature

BLEU	Bilingual Evaluation Understudy, an evaluation metric for MT
DWL	Discriminative Word Lexicon
LM	Language Model
MT	Machine Translation
PBMT	Phrase-Based Machine Translation
POS	Part-Of-Speech
RAP	Resolution of Anaphora Procedure
RBMT	Rule-Based Machine Translation
SDWL	Source DWL
SMT	Statistical Machine Translation
TM	Translation Model

Bibliography

[BHK07] S. Bangalore, P. Haffner, and S. Kanthak, "Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 152–159.

[BM00] S. Botley and T. Mcenery, "Discourse anaphora," in *Corpus-based and Computational Approaches to Discourse Anaphora*, ser. Studies in Corpus Linguistics, S. Botley and T. Mcenery, Eds. John Benjamins Publishing Company, 2000, ch. 1.

[BPPM93] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, Jun. 1993.

[Cha00] E. Charniak, "A Maximum-Entropy-Inspired Parser," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, ser. NAACL 2000. Association for Computational Linguistics, 2000, pp. 132–139.

[CJ05] E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Association for Computational Linguistics, 2005, pp. 173–180.

[Cry04] D. Crystal, *The Cambridge Encyclopedia of the English Language*, 2nd ed. Cambridge University Press, 2004.

[Dau04] H. Daumé III, "Notes on CG and LM-BFGS Optimization of Logistic Regression," 2004, paper available at http://www.umiacs.umd.edu/ hal/megam.

[DMR83] G. S. Dell, G. McKoon, and R. Ratcliff, "The Activation of Antecedent Information during the Processing of Anaphoric Reference in Reading," *Journal of Verbal Learning and Verbal Behavior*, vol. 22, no. 1, pp. 121–132, 1983.

[Gui12] L. Guillou, "Improving Pronoun Translation for Statistical Machine Translation," in *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL '12, Avignon, France, 2012.

[Har12] C. Hardmeier, "Discourse in Statistical Machine Translation. A Survey and A Case Study," *Discours. Revue de linguistique, psycholinguistique et informatique*, vol. 11, 2012.

[HF10] C. Hardmeier and M. Federico, "Modelling Pronominal Anaphora in Statistical Machine Translation," in *Proceedings of the International Workshop on Spoken Language Translation*, Paris, France, 2010.

[HNT12] C. Hardmeier, J. Nivre, and J. Tiedemann, "Document-Wide Decoding for Phrase-Based Statistical Machine Translation," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12. Association for Computational Linguistics, 2012, pp. 1179–1190.

[Hob78] J. R. Hobbs, "Resolving Pronoun References," *Lingua*, vol. 44, pp. 311–338, 1978.

[HS92] J. Hutchins and H. Somers, *An Introduction to Machine Translation*. Academic Press, 1992.

[HTS+11] C. Hardmeier, J. Tiedemann, M. Saers, M. Federico, and M. Prashant, "The Uppsala-FBK systems at WMT 2011," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, ser. WMT '11, San Francisco, USA, 2011.

[Jes54] O. Jespersen, *A modern English grammar on historical principles, part VII: Syntax*. Allen and Unwin, 1954.

[KH07] P. Koehn and H. Hoang, "Factored Translation Models," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2007, pp. 868–876.

[KL75] D. Klapholtz and A. Lockman, "Contextual Reference Resolution," *American Journal of Computational Linguistics*, 1975.

[Koe10] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2010.

[KOM03] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-based Translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, ser. NAACL '03. Association for Computational Linguistics, 2003, pp. 48–54.

[LK10] R. Le Nagard and P. Koehn, "Aiding Pronoun Translation with Co-Reference Resolution," in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, ser. WMT '10, Uppsala, Sweden, 2010.

[LL94] S. Lappin and H. Leass, "An Algorithm for Pronominal Anaphora Resolution," *Computational Linguistics*, vol. 20, no. 4, pp. 535–561, 1994.

[LW03] T. Liang and D.-S. Wu, "Automatic Pronominal Anaphora Resolution in English Texts," in *ROCLING*, 2003.

[MCN+11] M. Mediani, E. Cho, J. Niehues, T. Herrmann, and A. Waibel, "The KIT English-French Translation Systems for IWSLT 2011," ser. Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT), San Francisco, USA, Dec. 2011.

[MCS95] R. Mitkov, S.-k. Choi, and R. Sharp, "Anaphora Resolution in Machine Translation," in *Proceedings of the Sixth International conference on Theoretical and Methodological issues in Machine Translation*, Leuven, Belgium, 1995.

[MEO+12] R. Mitkov, R. Evans, C. Orăsan, I. Dornescu, and M. Rios, "Coreference Resolution: To What Extent Does It Help NLP Applications?" in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Springer Berlin Heidelberg, 2012, vol. 7499, pp. 16–27.

[MHN09] A. Mauser, S. Hasan, and H. Ney, "Extending statistical machine translation with discriminative and trigger-based lexicon models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '09. Singapore: Association for Computational Linguistics, 2009, pp. 210–218.

[MPB12] T. Meyer and A. Popescu-Belis, "Using Sense-Labeled Discourse Connectives for Statistical Machine Translation," in *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, ser. EACL 2012. Association for Computational Linguistics, 2012, pp. 129–138.

[MPBHG12] T. Meyer, A. Popescu-Belis, N. Hajlaoui, and A. Gesmundo, "Machine Translation of Labeled Discourse Connectives," in *Proceedings of AMTA 2012*, ser. AMTA 2012, 2012.

[Nic03] N. Nicolov, "Anaphora Resolution [Review of the book Anaphora Resolution. Ruslan Mitkov. Longman, 2002]," *IEEE Computational Intelligence Bulletin*, vol. 2, no. 1, pp. 31–32, June 2003.

[NNZ13] M. Novak, A. Nedoluzhko, and Z. Zabokrtsky, "Translation of "It" in a Deep Syntax Framework," in *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia, Bulgaria, 2013.

[Nov11] M. Novák, "Utilization of Anaphora in Machine Translation," in *WDS 2011 Proceedings of Contributed Papers: Part I*, ser. WDS 11, 2011, pp. 155–160.

[NW13] J. Niehues and A. Waibel, "An MT Error-Driven Discriminative Word Lexicon using Sentence Structure Features," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 512–520.

[NZM+12] J. Niehues, Y. Zhang, M. Mediani, T. Herrmann, E. Cho, and A. Waibel, "The Karlsruhe Institute of Technology Translation Systems for the

WMT 2012," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, 2012.

[Och03] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03. Association for Computational Linguistics, 2003, pp. 160–167.

[PBML+12] A. Popescu-Belis, T. Meyer, J. Liyanapathirana, B. Cartoni, and S. Zufferey, "Discourse-Level Annotation over Europarl for Machine Translation: Connectives and Pronouns," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, ser. LREC 2012, 2012, pp. 2716–2720.

[QKC04] L. Qiu, M.-Y. Kan, and T.-S. Chua, "A Public Reference Implementation of the RAP Anaphora Resolution Algorithm," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004.

[Sch95] H. Schmid, "Improvements In Part-of-Speech Tagging With an Application To German," in *Proceedings of the ACL SIGDAT-Workshop*, Dublin, 1995, pp. 47–50.

[SHTN13] S. Stymne, C. Hardmeier, J. Tiedemann, and J. Nivre, "Feature Weight Optimization for Discourse-Level SMT," in *Proceedings of the Workshop on Discourse in Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 60–69.

[SL08] H. Schmid and F. Laws, "Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging," in *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, ser. COLING '08, Manchester, UK, 2008.

[SNL01] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A Machine Learning Approach to Coreference Resolution of Noun Phrases," *Computational Linguistics*, vol. 27, no. 4, pp. 521–544, 2001.

[TB01] E. Tognini-Bonelli, *Corpus Linguistics at Work*, ser. Studies in Corpus Linguistics. John Benjamins, 2001, vol. 6.

[vDK00] K. van Deemter and R. Kibble, "On Coreferring: Coreference in MUC and Related Annotation Schemes," *Computational Linguistics*, vol. 26, no. 4, pp. 629–637, 2000.