Improving Pronoun Translation for Statistical Machine Translation

Liane Guillou
School of Informatics
University of Edinburgh
Edinburgh, UK, EH8 9AB
L.K.Guillou@sms.ed.ac.uk

Abstract

Machine Translation is a well–established field, yet the majority of current systems translate sentences in isolation, losing valuable contextual information from previously translated sentences in the discourse. One important type of contextual information concerns who or what a coreferring pronoun corefers to (i.e., its *antecedent*). Languages differ significantly in how they achieve coreference, and awareness of antecedents is important in choosing the correct pronoun. Disregarding a pronoun's antecedent in translation can lead to inappropriate coreferring forms in the target text, seriously degrading a reader's ability to understand it.

This work assesses the extent to which *source-language annotation* of coreferring pronouns can improve English–Czech Statistical Machine Translation (SMT). As with previous attempts that use this method, the results show little improvement. This paper attempts to explain why and to provide insight into the factors affecting performance.

1 Introduction

It is well-known that in many natural languages, a pronoun that corefers must bear similar features to its antecedent. These can include similar number, gender (morphological or referential), and/or animacy. If a pronoun and its antecedent occur in the same unit of translation (N-gram or syntactic tree), these agreement features can influence the translation. But this locality cannot be guaranteed in either phrase-based or syntax-based Statistical Machine Translation (SMT). If it is not within the same unit, a coreferring pronoun will be translated without knowledge of its antecedent, meaning that its translation will simply reflect local frequency. Incorrectly translating a pronoun can result in readers/listeners identifying the wrong antecedent, which can mislead or confuse them.

There have been two recent attempts to solve this problem within the framework of phrase-based SMT (Hardmeier & Federico, 2010; Le Nagard & Koehn, 2010). Both involve *annotation projection*, which in this context means annotating coreferential pronouns in the source-language with features derived from the translation of their aligned antecedents, and then building a *translation model* of the annotated forms. When translating a coreferring pronoun in a new source-language text, the antecedent is identified and its translation used (differently in the two attempts cited above) to annotate the pronoun prior to translation.

The aim of this work was to better understand why neither of the previous attempts achieved more than a small improvement in translation quality associated with coreferring pronouns. Only by understanding this will it be possible to ascertain whether the method of *annotation projection* is intrinsically flawed or the unexpectedly small improvement is due to other factors.

Errors can arise when:

1. Deciding whether or not a third person pronoun corefers;
2. Identifying the pronoun antecedent;
3. Identifying the head of the antecedent, which serves as the source of its features;
4. Aligning the source and target texts at the phrase and word levels.

Proceedings of the EACL 2012 Student Research Workshop, pages 1–10,
Avignon, France, 26 April 2012.

Factoring out the first two decisions would show whether the lack of significant improvement was simply due to imperfect coreference resolution. In order to control for these errors several different manually annotated versions of the Penn *Wall Street Journal* corpus were used, each providing different annotations over the same text. The BBN Pronoun Coreference and Entity Type corpus (Weischedel & Brunstein, 2005) was used to provide coreference information in the source-language and exclude non-referential pronouns. It also formed the source-language side of the parallel training corpus. The PCEDT 2.0 corpus (Hajič et al., 2011), which contains a close Czech translation of the Penn *Wall Street Journal* corpus, provided reference translations for testing and the target-language side of the parallel corpus for training. To minimise (although not completely eliminate) errors associated with antecedent head identification (item 3 above), the parse trees in the Penn Treebank 3.0 corpus (Marcus et al., 1999) were used. The *gold standard* annotation provided by these corpora allowed me to assume perfect identification of coreferring pronouns and coreference resolution and near–perfect antecedent head noun identification. These assumptions could not be made if state-of-the-art methods had been used as they cannot yet achieve sufficiently high levels of accuracy.

The remainder of the paper is structured as follows. The use of pronominal coreference in English and Czech and the problem of anaphora resolution are described in Section 2. The works of Le Nagard & Koehn (2010) and Hardmeier & Federico (2010) are discussed in Section 3, and the source-language annotation projection method is described in Section 4. The results are presented and discussed in Section 5 and future work is outlined in Section 6.

2 Background

2.1 Anaphora Resolution

Anaphora resolution involves identifying the antecedent of a referring expression, typically a pronoun or noun phrase that is used to refer to something previously mentioned in the discourse (its antecedent). Where multiple referring expressions refer to the same antecedent, they are said to be *coreferential*. Anaphora resolution and the related task of coreference resolution have been the subject of considerable research within Natural Language Processing (NLP). Excellent surveys are provided by Strube (2007) and Ng (2010).

Unresolved anaphora can add significant translation ambiguity, and their incorrect translation can significantly decrease a reader's ability to understand a text. Accurate coreference in translation is therefore necessary in order to produce understandable and cohesive texts. This justifies recent interest (Le Nagard & Koehn, 2010; Hardmeier & Federico, 2010) and motivates the work presented in this paper.

2.2 Pronominal Coreference in English

Whilst English makes some use of case, it lacks the grammatical gender found in other languages. For monolingual speakers, the relatively few different pronoun forms in English make sentences easy to generate: Pronoun choice depends on the number and gender of the entity to which they refer. For example, when talking about ownership of a book, English uses the pronouns "his/her" to refer to a book that belongs to a male/female owner, and "their" to refer to one with multiple owners (irrespective of their gender). One source of difficulty is that the pronoun "it" has both a coreferential and a pleonastic function. A pleonastic pronoun is one that is not referential. For example, in the sentence "It is raining", "it" does not corefer with anything. Coreference resolution algorithms must exclude such instances in order to prevent the erroneous identification of an antecedent when one does not exist.

2.3 Pronominal Coreference in Czech

Czech, like other Slavic languages, is highly inflective. It is also a free word order language, in which word order reflects the information structure of the sentence within the current discourse. Czech has seven cases and four grammatical genders: masculine animate (for people and animals), masculine inanimate (for inanimate objects), feminine and neuter. (With feminine and neuter genders, animacy is not grammatically marked.) In Czech, a pronoun must agree in number, gender and animacy with its antecedent. The morphological form of possessive pronouns depends not only on the possessor but also the object in possession. Moreover, reflexive pronouns (both personal and possessive) are commonly used. In addition, Czech is a pro-drop language, whereby an

explicit subject pronoun may be omitted if it is inferable from other grammatical features such as verb morphology. This is in contrast with English which exhibits relatively fixed Subject-Verb-Object (SVO) order and only drops subject pronouns in imperatives (e.g. "Stop babbling") and coordinated VPs.

Differences between the choice of coreferring expressions used in English and Czech can be seen in the following simple examples:

1. The dog has a ball. I can see **it** playing outside.
2. The cow is in the field. I can see **it** grazing.
3. The car is in the garage. I will take **it** to work.

In each example, the English pronoun "it" refers to an entity that has a different gender in Czech. Its correct translation requires identifying the gender (and number) of its antecedent and ensuring that the pronoun agrees. In 1 "it" refers to the dog ("pes", masculine, animate) and should be translated as "ho". In 2, "it" refers to the cow ("kráva", feminine) and should be translated as "ji". In 3, "it" refers to the car ("auto", neuter) and should be translated as "ho".

In some cases, the same pronoun is used for both animate and inanimate masculine genders, but in general, different pronouns are used. For example, with possessive reflexive pronouns in the accusative case:

English: *I admired* ***my (own)*** *dog*
Czech: *Obdivoval jsme* ***svého*** *psa*

English: *I admired* ***my (own)*** *castle*
Czech: *Obdivoval jsme* ***svůj*** *hrad*

Here "svého" is used to refer to a dog (masculine animate, singular) and "svůj" to refer to a castle (masculine inanimate, singular), both of which belong to the speaker.

Because a pronoun may take a large number of morphological forms in Czech and because case is not checked in annotation projection, the method presented here for translating coreferring pronouns does not guarantee their correct form.

3 Related Work

Early work on integrating anaphora resolution with Machine Translation includes the rule-based approaches of Mitkov et al. (1995) and Lappin & Leass (1994) and the transfer-based approach of Saggion & Carvalho (1994). Work in the 1990's culminated in the publication of a special issue of *Machine Translation* on anaphora resolution (Mitkov, 1999). Work then appears to have been on hold until papers were published by Le Nagard & Koehn (2010) and Hardmeier & Federico (2010). This resurgence of interest follows advances since the 1990's which have made new approaches possible.

The work described in this paper resembles that of Le Nagard & Koehn (2010), with two main differences. The first is the use of manually annotated corpora to extract coreference information and morphological properties of the target translations of the antecedents. The second lies in the choice of language pair. They consider English-French translation, focussing on gender-correct translation of the third person pronouns "it" and "they". Coreference is more complex in Czech with both number and gender influencing pronoun selection. Annotating pronouns with both number and gender further exacerbates the problem of data sparseness in the training data, but this cannot be avoided if the aim is to improve their translation. This work also accommodates a wider range of English pronouns.

In contrast, Hardmeier & Federico (2010) focus on English-German translation and model coreference using a word dependency module integrated within the log-linear SMT model as an additional feature function.

Annotation projection has been used elsewhere in SMT. Gimpel & Smith (2008) use it to capture long–distance phenomena within a single sentence in the source-language text via the extraction of sentence-level contextual features, which are used to augment SMT translation models and better predict phrase translation. Projection techniques have also been applied to multilingual Word Sense Disambiguation whereby the sense of a word may be determined in another language (Diab, 2004; Khapra et al., 2009).

4 Methodology

4.1 Overview

I have followed Le Nagard & Koehn (2010) in using a two-step approach to translation, with *annotation projection* incorporated as a pre-processing

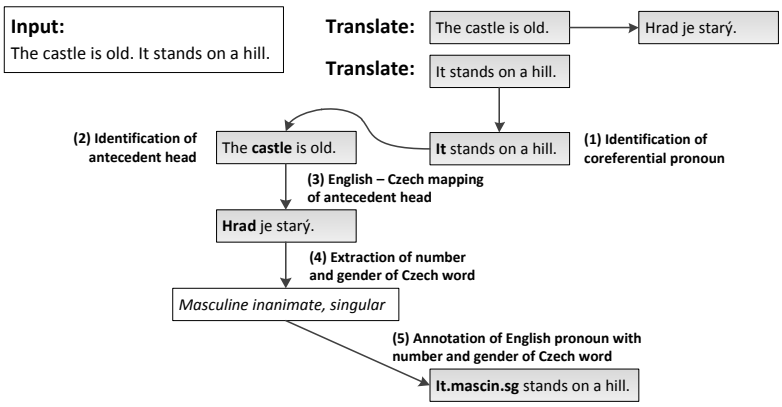

Figure 1: Overview of the Annotation Process

task. In the first step, pronouns are annotated in the source-language text before the text is translated by a phrase-based SMT system in the second step. This approach leaves the translation process unaffected. In this work, the following pronouns are annotated: third person personal pronouns (except instances of "it" that are pleonastic or that corefer with clauses or VPs), reflexive personal pronouns and possessive pronouns, including reflexive possessives. Relative pronouns are excluded as they are local dependencies in both English and Czech and this work is concerned with the longer range dependencies typically exhibited by the previously listed pronoun types.

Annotation of the English source-language text and its subsequent translation into Czech is achieved using two phrase-based translation systems. The first, hereafter called the *Baseline system*, is trained using English and Czech sentence–aligned parallel training data with no annotation. The second system, hereafter called the *Annotated system*, is trained using the same target data, but in the source-language text, each coreferring pronoun has been annotated with number, gender and animacy features. These are obtained from the existing (Czech reference) translation of the head of its English antecedent. Word alignment of English and Czech is obtained from the PCEDT 2.0 alignment file which maps English words to their corresponding t-Layer (deep syntactic, tectogrammatical) node in the Czech translation. Starting with this t-Layer node the annotation layers of the PCEDT 2.0 corpus are traversed and the number and gender of the Czech word are extracted from the morphological layer (m-Layer).

The Baseline system serves a dual purpose. It forms the first stage of the two-step translation process, and as described in Section 5, it provides a baseline against which Annotated system translations are compared.

The annotation process used here is shown in Figure 1. It identifies coreferential pronouns and their antecedents using the annotation in the BBN Pronoun Coreference and Entity Type corpus, and obtains the Czech translation of the English antecedent from the translation produced by the Baseline system. Because many antecedents come from previous sentences, these sentences must be translated before translating the current sentence. Here I follow Le Nagard & Koehn (2010) in translating the complete source-language text using the Baseline system and then extracting the (here, Czech) translations of the English antecedents from the output. This provides a simple solution to the problem of obtaining the Czech translation prior to annotation. In contrast Hardmeier & Federico (2010) translate sentence by sentence using a process which was deemed to be more complex than was necessary for this project.

The English text is annotated such that all coreferential pronouns whose antecedents have an identifiable Czech translation are marked with the number and gender of that Czech word. The output of the annotation process is thus the same English text that was input to the Baseline system, with the addition of annotation of the coreferential pronouns. This annotated English text is then translated using the Annotated translation system, the output of which is the final translation.

	Training	Dev.	Final
Parallel Sentences	47,549	280	540
Czech Words	955,018	5,467	10,110
English Words	1,024,438	6,114	11,907

Table 1: Sizes of the training and testing datasets

4.2 Baseline and Annotated systems

Both systems are phrase-based SMT models, trained using the Moses toolkit (Hoang et al., 2007). They share the same 3-gram language model constructed from the target-side text of the parallel training corpus and the Czech monolingual 2010 and 2011 News Crawl corpora[1]. The language model was constructed using the SRILM toolkit (Stolcke, 2002) with interpolated Kneser-Ney discounting (Kneser & Ney, 1995). In addition, both systems are forced to use the same word alignments (constructed using Giza++ (Och & Ney, 2003) in both language pair directions and using *stemmed* training data in which words are limited to the first four characters) in order to mitigate the effects of Czech word inflection on word alignment statistics. This helps to ensure that the Czech translation of the head of the antecedent remains constant in both steps of the two-step process. If this were to change it would defeat the purpose of pronoun annotation as different Czech translations could result in different gender and/or number.

The Baseline system was trained using the Penn *Wall Street Journal* corpus with no annotation, while the Annotated system was trained with an annotated version of the same text (see Table 1), with the target-language text being the same in both cases. The Penn *Wall Street Journal* corpus was annotated using the process described above, with the number and gender of the Czech translation of the antecedent head obtained from the PCEDT 2.0 alignment file.

4.3 Processing test files

Two test files were used (see Table 1) – one called 'Final' and the other, 'Development' (Dev). A test file is first translated using the Baseline system with a trace added to the Moses decoder. Each coreferential English pronoun is then identified using the BBN Pronoun Coreference and Entity Type corpus and the head of its antecedent is extracted from the annotated NPs in the Penn Treebank 3.0 corpus. The sentence number and word position of the English pronoun and its antecedent head noun(s) are extracted from the input English text and used to identify the English/Czech phrase pairs that contain the Czech translations of the English words. Using this information together with the phrase alignments (output by the Moses decoder) and the phrase-internal word alignments in the phrase translation table, a Czech translation is obtained from the Baseline system. Number, gender and animacy (if masculine) features of the Czech word identified as the translation of the head of the antecedent are extracted from a pre-built morphological dictionary of Czech words constructed from the PCEDT 2.0 corpus for the purpose of this work. A copy of the original English test file is then constructed, with each coreferential pronoun annotated with the extracted Czech features.

The design of this process reflects two assumptions. First, the annotation of the Czech words in the m-Layer of the PCEDT 2.0 corpus is both accurate and consistent. Second, as the Baseline and Annotated systems were trained using the same word alignments, the Czech translation of the head of the English antecedent should be the same in the output of both. Judging by the very small number of cases in which the antecedent translations differed (3 out of 458 instances), this assumption was proved to be reasonable. These differences were due to the use of different phrase tables for each system as a result of training on different data (i.e. the annotation of English pronouns or lack thereof). This would not be an issue for single-step translation systems such as that used by Hardmeier & Federico (2010).

4.4 Evaluation

No standard method yet exists for evaluating pronoun translation in SMT. Early work focussed on the development of techniques for anaphora resolution and their integration within Machine Translation (Lappin & Leass, 1994; Saggion & Carvalho, 1994; Mitkov et al., 1995), with little mention of evaluation. In recent work, evaluation has become much more important. Both Le Nagard & Koehn (2010) and Hardmeier & Federico (2010) consider and reject BLEU (Papineni et al., 2002) as ill-suited for evaluating pronoun translation. While Hardmeier & Federico propose and

[1] Provided for the Sixth EMNLP Workshop on Statistical Machine Translation (Callison-Burch et al., 2011)

use a strict recall and precision based metric for English–German translation, I found it unsuitable for English–Czech translation, given the highly inflective nature of Czech.

Given the importance of evaluation to the goal of assessing the effectiveness of *annotation projection* for improving the translation of coreferring pronouns, I carried out two separate types of evaluation — an automated evaluation which could be applied to the entire test set, and an in-depth manual assessment that might provide more information, but could only be performed on a subset of the test set. The automated evaluation is based on the fact that a Czech pronoun must agree in number and gender with its antecedent. Thus one can count the number of pronouns in the translation output for which this agreement holds, rather than simply score the output against a single reference translation. To obtain these figures, the automated evaluation process counted:

1. Total pronouns in the input English test file.
2. Total English pronouns identified as coreferential, as per the annotation of the BBN Pronoun Coreference and Entity Type corpus.
3. Total coreferential English pronouns that are annotated by the annotation process.
4. Total coreferential English pronouns that are aligned with any Czech translation.
5. Total coreferential English pronouns translated as any Czech pronoun.
6. Total coreferential English pronouns translated as a Czech pronoun corresponding to a **valid** translation of the English pronoun.
7. Total coreferential English pronouns translated as a Czech pronoun (that is a valid translation of the English pronoun) agreeing in number and gender with the antecedent.

The representation of valid Czech translations of English pronouns takes the form of a list provided by an expert in Czech NLP, which ignores case and focusses solely on number and gender.

In contrast, the manual evaluation carried out by that same expert, who is also a native speaker of Czech, was used to determine whether deviations from the single reference translation provided in the PCEDT 2.0 corpus were valid alternatives or simply poor translations. The following judgements were provided:

1. Whether the pronoun had been translated correctly, or in the case of a dropped pronoun, whether pro-drop was appropriate;
2. If the pronoun translation was incorrect, whether a native Czech speaker would still be able to derive the meaning;
3. For input to the Annotated system, whether the pronoun had been correctly annotated with respect to the Czech translation of its identified antecedent;
4. Where an English pronoun was translated differently by the Baseline and Annotated systems, which was better. If both translated an English pronoun to a valid Czech translation, equal correctness was assumed.

In order to ensure that the manual assessor was directed to the Czech translations aligned to the English pronouns, additional markup was automatically inserted into the English and Czech texts: (1) coreferential pronouns in both English and Czech texts were marked with the head noun of their antecedent (denoted by *), and (2) coreferential pronouns in the English source texts were marked with the Czech translation of the antecedent head, and those in the Czech target texts were marked with the original English pronoun that they were aligned to:

English text input to the Baseline system: *the u.s. , claiming some success in its trade diplomacy , ...*

Czech translation output by the Baseline system: *usa , tvrdí někteří její**(its)** obchodní úspěch v diplomacii , ...*

English text input to the Annotated system: *the u.s.* , claiming some success in its**(u.s.,usa)**.mascin.pl trade diplomacy , ...*

Czech translation output by the Annotated system: *usa ,* tvrdí někteří úspěchu ve své**(its.mascin.pl)** obchodní diplomacii , ...*

5 Results and Discussion

5.1 Automated Evaluation

Automated evaluation of both "Development" and "Final" test sets (see Table 2) shows that even factoring out the problems of accurate identification of coreferring pronouns, coreference resolution and antecedent head–finding, does not improve performance of the Annotated system much above that of the Baseline.

	Dev.		Final	
	Baseline	Annotated	Baseline	Annotated
Total pronouns in English file	156	156	350	350
Total pronouns identified as coreferential	141	141	331	331
Annotated coreferential English pronouns	–	117	–	278
Coreferential English pronouns aligned with any Czech translation	141	141	317	317
Coreferential English pronouns translated as Czech pronouns	71	75	198	198
Czech pronouns that are valid translations of the English pronouns	63	71	182	182
Czech pronouns that are valid translations of the English pronouns and that match their antecedent in number and gender	44	46	142	146

Table 2: Automated Evaluation Results for both test sets

Criterion	Baseline System Better	Annotated System Better	Systems Equal
Overall quality	**9/31** (29.03%)	**11/31** (35.48%)	**11/31** (35.48%)
Quality when annotation is correct	**3/18** (16.67%)	**9/18** (50.00%)	**6/18** (33.33%)

Table 3: Manual Evaluation Results: A direct comparison of pronoun translations that differ between systems

Taking the accuracy of pronoun translation to be the proportion of coreferential English pronouns having a valid Czech translation that agrees in both number and gender with their antecedent, yields the following on the two test sets:

Baseline system:

Development — 44/141 (31.21%)

Final — 142/331 (42.90%)

Annotated system:

Development — 46/141 (32.62%)

Final — 146/331 (44.10%)

There are, however, several reasons for not taking this evaluation as definitive. Firstly, it relies on the accuracy of the word alignments output by the decoder to identify the Czech translations of the English pronoun and its antecedent. Secondly, these results fail to capture variation between the translations produced by the Baseline and Annotated systems. Whilst there is a fairly high degree of overlap, for approximately 1/3 of the "Development" set pronouns and 1/6 of the "Final" set pronouns, the Czech translation is different. Since the goal of this work was *to understand what is needed in order to improve the translation of coreferential pronouns*, manual evaluation was critical for understanding the potential capabilities of source-side annotation.

5.2 Manual Evaluation

The sample files provided for manual evaluation contained 31 pronouns for which the translations provided by the two systems differed (*differences*) and 72 for which the translation provided by the systems was the same (*matches*). Thus, the sample comprised 103 of the 472 coreferential pronouns (about 22%) from across both test sets. Of this sample, it is the *differences* that indicate the relative performance of the two systems. Of the 31 pronouns in this set, 16 were 3^{rd}-person pronouns, 2 were reflexive personal pronouns and 13 were possessive pronouns.

The results corresponding to evaluation criterion 4 in Section 4.4 provide a comparison of the overall quality of pronoun translation for both systems. These results for the "Development" and "Final" test sets (see Table 3) suggest that the performance of the Annotated system is comparable with, and even marginally better than, that of the Baseline system, especially when the pronoun annotation is correct.

An example of where the Annotated system produces a better translation than the Baseline system is:

Annotated English: *he said **mexico** could be one of the next countries to be removed from the priority list because of **its.neut.sg** efforts to craft a new patent law .*

Baseline translation: *řekl , že **mexiko** by mohl být jeden z dalších zemí , aby byl odvolán z prioritou seznam , protože **její** snahy podpořit nové patentový zákon .*

Annotated translation: *řekl , že **mexiko** by mohl být jeden z dalších zemí , aby byl odvolán z prioritou seznam , protože **jeho** snahy podpořit nové patentový zákon .*

In this example, the English pronoun "its", which refers to "mexico" is annotated as neuter and singular (as extracted from the Baseline translation). Both systems translate "mexico" as "mexiko" (neuter, singular) but differ in their translation of the pronoun. The Baseline system translates "its" incorrectly as "její" (feminine, singular), whereas the Annotated system produces

the more correct translation: “jeho” (neuter, singular), which agrees with the antecedent in both number and gender.

An analysis of the judgements on the remaining three evaluation criteria (outlined in Section 4.4) for the 31 *differences* provides further information. The Baseline system appears to be more accurate, with 19 pronouns either correctly translated (in terms of number and gender) or appropriately dropped, compared with 17 for the Annotated system. Of those pronouns, the meaning could still be understood for 7/12 for the Baseline system compared with 8/14 for the Annotated system. On the surface this may seem strange but it appears to be due to a small number of cases in which the translations produced by both systems were incorrect but those produced by the Annotated system were deemed to be marginally better. Due to the small sample size it is difficult to form a complete picture of where one system may perform consistently better than the other. The annotation of both number and gender was accurate for 18 pronouns. Whilst this accuracy is not particularly high, the results (see Table 3) suggest that translation is more accurate for those pronouns that are correctly annotated.

Whilst pro-drop in Czech was not explicitly handled in the annotation process, manual evaluation revealed that both systems were able to successfully ‘learn’ a few (local) scenarios in which pro-drop is appropriate. This was unexpected but found to be due to instances in which there are short distances between the pronoun and verb in English. For example, many of the occurrences of “she” in English appear in the context of “she said...” and are translated correctly with the verb form “...řekla...”.

An example of where the Annotated system correctly drops a pronoun is:

Annotated English: *“ this is the worst **shakeout** ever in the junk market , and it could take years before **it.fem.sg** ’ s over , ” says mark bachmann , a senior vice president at standard & poor ’ s corp . , a credit rating company .*

Baseline translation: *“ je to nejhorší **krize** , kdy na trhu s rizikovými obligacemi , a to může trvat roky , než je **to** pryč , ” říká mark bachmann , hlavní viceprezident společnosti standard & poor ’s corp . , úvěrový rating společnosti .*

Annotated translation: *“ je to nejhorší **krize** , kdy na trhu s rizikovými obligacemi , a to může trvat roky , než je **!!** pryč , ” říká mark bachmann , hlavní viceprezident společnosti standard & poor ’s corp . , úvěrový rating společnosti .*

In this example, the Baseline system translates “it” incorrectly as the neuter singular pronoun “to”, whereas the Annotated system correctly drops the subject pronoun (indicated by **!!**) — this is a less trivial example than “she said”. In the case of the Baseline translation “to” could be interpreted as referring to the whole event, which would be correct, but poor from a stylistic point of view.

An example of where the Annotated system fails to drop a pronoun is:

Annotated English: ***taiwan** has improved **its.mascin.sg*** standing with the u.s. by initialing a bilateral copyright agreement , amending **its.mascin.sg**** trademark law and introducing legislation to protect foreign movie producers from unauthorized showings of their.mascan.pl films .*

Annotated translation: ***tchaj-wan** zlepšení **své** postavení s usa o initialing bilaterálních autorských práv na **jeho** obchodní dohody , úprava zákona a zavedení zákona na ochranu zahraniční filmové producenty z neoprávněné showings svých filmů .*

Reference translation: ***tchaj-wan** zlepšil svou reputaci v usa , když podepsal bilaterální smlouvu o autorských právech , pozměnil **!!** zákon o ochranných známkách a zavedl legislativu na ochranu zahraničních filmových producentů proti neautorizovanému promítání jejich filmů .*

In this example, the English pronoun “its”, which refers to “taiwan” is annotated as masculine inanimate and singular. The first occurrence of “its” is marked by * and the second occurrence by ** in the annotated English text above. The second occurrence should be translated either as a reflexive pronoun (as the first occurrence is correctly translated) or it should be dropped as in the reference translation (**!!** indicates the position of the dropped pronoun).

In addition to the judgements, the manual assessor also provided feedback on the evaluation task. One of the major difficulties encountered concerned the translation of pronouns in sentences which exhibit poor syntactic structure. This is a criticism of Machine Translation as a whole, but of the manual evaluation of pronoun translation in particular, since the choice of coreferring form is sensitive to syntactic structure. Also the effects of poor syntactic structure are likely to introduce an additional element of subjectivity if the assessor must first interpret the structure of the sentences output by the translation systems.

5.3 Potential Sources of Error

Related errors that may have contributed to the Annotated system not providing a significant improvement over the Baseline include: (1) incor-

rect identification of the English antecedent head noun, (2) incorrect identification of the Czech translation of the antecedent head noun in the Baseline output due to errors in the word alignments, and (3) errors in the PCEDT 2.0 alignment file (affecting training only). While "perfect" annotation of the BBN Pronoun Coreference and Entity Type, the PCEDT 2.0 and the Penn Treebank 3.0 corpora has been assumed, errors in these corpora cannot be completely ruled out.

6 Conclusion and Future Work

Despite factoring out three major sources of error — identifying coreferential pronouns, finding their antecedents, and identifying the head of each antecedent — through the use of manually annotated corpora, the results of the Annotated system show only a small improvement over the Baseline system. Two possible reasons for this are that the statistics in the phrase translation table have been weakened in the Annotated system as a result of including both number and gender in the annotation and that the size of the training corpus is relatively small.

However, more significant may be the availability of only a single reference translation. This affects the development and application of automated evaluation metrics as a single reference cannot capture the variety of possible valid translations. Coreference can be achieved without explicit pronouns. This is true of both English and Czech, with sentences that contain pronouns having common paraphrases that lack them. For example,

the u.s. , claiming some success in ***its*** *trade diplomacy , ...*

can be paraphrased as:

the u.s. , claiming some success in trade diplomacy , ...

A target-language translation of the former might actually be a translation of the latter, and hence lack the pronoun shown in bold. Given the range of variability in whether pronouns are used in conveying coreference, the availability of only a single reference translation is a real problem.

Improving the accuracy of coreferential pronoun translation remains an open problem in Machine Translation and as such there is great scope for future work in this area. The investigation reported here suggests that it is not sufficient to focus solely on the source-side and further operations on the target side (besides post-translation application of a target-language model) need also be considered. Other target–side operations could involve the extraction of features to score multiple candidate translations in the selection of the 'best' option – for example, to 'learn' scenarios in which pro-drop is appropriate and to select translations that contain pronouns of the correct morphological inflection. This requires identification of features in the target side, their extraction and incorporation in the translation process which could be difficult to achieve within a purely statistical framework given that the antecedent of a pronoun may be arbitrarily distant in the previous discourse.

The aim of this work was to better understand why previous attempts at using annotation projection in pronoun translation showed less than expected improvement. Thus it would be beneficial to conduct an error analysis to show the frequency of the errors described in Section 5.3 appear.

I will also be exploring other directions related to problems identified during the course of the work completed to date. These include, but are not limited to, handling pronoun dropping in pro-drop languages, developing pronoun-specific automated evaluation metrics and addressing the problem of having only one reference translation for use with such metrics. In this regard, I will be considering the use of paraphrase techniques to generate synthetic reference translations to augment an existing reference translation set. Initial efforts will focus on adapting the approach of Kauchak & Barzilay (2006) and back–translation methods for extracting paraphrases (Bannard & Callison-Burch, 2005) to the more specific problem of pronoun variation.

Acknowledgements

I would like to thank Bonnie Webber (University of Edinburgh) who supervised this project and Markéta Lopatková (Charles University) who provided the much needed Czech language assistance. I am very grateful to Ondřej Bojar (Charles University) for his numerous helpful suggestions and to the Institute of Formal and Applied Linguistics (Charles University) for providing the PCEDT 2.0 corpus. I would also like to thank Wolodja Wentland and the three anonymous reviewers for their feedback.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 597–604.

Chris Callison-Burch, Philipp Koehn, Christof Monz and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64.

Mona Diab. 2004. An Unsupervised Approach for Bootstrapping Arabic Sense Tagging. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 43–50.

Kevin Gimpel and Noah A. Smith. 2008. Rich Source-Side Context for Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 9–17.

Barbara J. Grosz, Scott Weinstein and Aravind K. Joshi. 1995. Centering: A Framework for Modeling the Local Coherence Of Discourse. *Computational Linguistics*, 21(2):203–225.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, pages 283–290.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens. Chris Dyer, Ondřej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Jerry R. Hobbs. 1978. Resolving Pronominal References. *Lingua*, 44:311–338.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová and Zdeněk Žabokrtský. 2011. Prague Czech-English Dependency Treebank 2.0. Institute of Formal and Applied Linguistics. Prague, Czech Republic.

David Kauchak and Regina Barzilay. 2006. Paraphrasing For Automatic Evaluation. In *Proceedings of the Main Conference on Human Language Technology Conference of the NAACL*, June 5–7, New York, USA, pages 455–462.

Mitesh M. Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya. 2009. Projecting Parameters for Multilingual Word Sense Disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, August 6–7, Singapore, pages 459–467.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-Off for M-gram Language Modeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 9–12, Detroit, USA, 1:181–184.

Shalom Lappin and Herbert J. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20:535–561.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding Pronoun Translation with Co-reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261.

Vincent Ng. 2010. Supervised Noun Phrase Coreference Research: The first 15 years. In *Proceedings of the 48th Meeting of the ACL*, pages 1396–1411.

Mitchell P. Marcus, Beatrice Santorini, Mary A. Marcinkiewicz and Ann Taylor. 1999. Penn Treebank 3.0 LDC Calalog No.: LDC99T42. Linguistic Data Consortium.

Ruslan Mitkov, Sung-Kwon Choi and Randall Sharp. 1995. Anaphora Resolution in Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, July 5-7, Leuven, Belgium, pages 5–7.

Ruslan Mitkov. 1999. Introduction: Special Issue on Anaphora Resolution in Machine Translation and Multilingual NLP. *Machine Translation*, 14:159–161.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318.

Horacio Saggion and Ariadne Carvalho. 1994. Anaphora Resolution in a Machine Translation System. In *Proceedings of the International Conference on Machine Translation: Ten Years On*, November, Cranfield, UK, 4.1-4.14.

Andreas Stolcke. 2002. SRILM — An Extensible Language Modeling Toolkit. In *Proceedings of International Conference on Spoken Language Processing*, September 16-20, Denver, USA, 2:901–904.

Michael Strube. 2007. Corpus-based and Machine Learning Approaches to Anaphora Resolution. *Anaphors in Text: Cognitive, Formal and Applied Approaches to Anaphoric Reference*, John Benjamins Pub Co.

Ralph Weischedel and Ada Brunstein. 2005. BBN Pronoun Coreference and Entity Type Corpus LDC Calalog No.: LDC2005T33. Linguistic Data Consortium.