Modelling Pronominal Anaphora in Statistical Machine Translation

Christian Hardmeier and Marcello Federico

Fondazione Bruno Kessler
Human Language Technologies
Via Sommarive, 18
38123 Trento, Italy
{hardmeier,federico}@fbk.eu

Abstract

Current Statistical Machine Translation (SMT) systems translate texts sentence by sentence without considering any cross-sentential context. Assuming independence between sentences makes it difficult to take certain translation decisions when the necessary information cannot be determined locally. We argue for the necessity to include cross-sentence dependencies in SMT. As a case in point, we study the problem of pronominal anaphora translation by manually evaluating German-English SMT output. We then present a word dependency model for SMT, which can represent links between word pairs in the same or in different sentences. We use this model to integrate the output of a coreference resolution system into English-German SMT with a view to improving the translation of anaphoric pronouns.

1. Introduction

Statistical Machine Translation (SMT) is strongly focused on local phenomena. The first models of the modern SMT paradigm published in the early 1990s [1, 2] impose strong independence assumptions on the words in a sentence and only take into account a very limited context consisting of the one or two immediately preceding words in the target language for each word the system outputs. Phrase-based and syntax-based SMT, the two currently dominant paradigms, relax these independence assumptions by considering more local dependencies, also in the source language. In syntax-based Machine Translation (MT), some long-range dependencies inside the sentence can be accommodated. However, even advanced MT systems still assume that texts can be translated sentence by sentence and that the sentences in a text are strictly independent of one another.

In this paper, we focus on the translation of pronouns in SMT. Pronominal reference is a discourse-level phenomenon which frequently occurs in many text genres and which cannot be handled satisfactorily under the assumption that sentences are mutually independent. We present the results of a case study about how an existing MT system copes with pronoun translation and demonstrate that pronoun choice is a problem for current SMT systems and go on to propose a model for integrating long-range dependencies between word pairs, optionally crossing sentence boundaries, into SMT. Finally, we describe a method to evaluate pronoun translation automatically and present some experimental results.

We argue that SMT research has reached a point where it is useful to start thinking about what is beyond the next sentence boundary. By presenting a model that is capable of representing links between words in different sentences, as well as between remote words in the same sentence, we hope to open a way towards making texts connected by integrating more abstract, non-local information into SMT.

2. Pronominal Anaphora in MT

Pronominal anaphora is the use of pronominal expressions to refer to "something previously mentioned in the discourse" [3], as in the following example, where *them* in the second sentence refers back to *the Catholics* in the first:

> *The Catholics* described the situation as "safe" and "protecting." This made *them* "relaxed and peaceful."

Anaphora is a very common phenomenon found in almost all kinds of texts. The reference link can be local to the sentence, or it can cross sentence boundaries. In the first case, it may be handled correctly by the local dependencies of the SMT language model, but this becomes increasingly unlikely as the distance between the referring pronoun and its antecedent increases. The second, non-local case is not covered by current SMT models at all.

It is worth pointing out that a pronoun may well be translated correctly even without the benefit of a specific anaphora model. In the example cited above, the plural pronoun *them* would probably be rendered as *sie* by a naïve English-German SMT system, which is very likely to be a good choice. When translating into a language with gender-marked plural pronouns, however, pronoun choice might be more difficult.

To show that pronominal anaphora is indeed a problem for current SMT, we studied the performance of one of our SMT systems on personal pronouns. The sample examined

Document		masc. sg.	fem. sg.	neuter sg.	plural	polite address	reflexive	demonstrative	pron. + prep.	total	
1	aktualne.cz	1/ 1	–/ 1	–/ 1	–/ 1	–/ –	–/ 2	1/ 1	–/ 2	2/ 9	22 %
2	spiegel	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–
3	bbc	5/ 8	6/23	1/ 2	–/ –	1/ 4	–/ 4	2/ 2	–/ –	15/ 43	35 %
4	bbc	9/11	1/ 2	2/ 2	–/ –	–/ –	–/ –	–/ –	1/ 1	13/ 16	81 %
5	times-of-london	1/ 3	2/ 2	–/ –	7/10	–/ –	–/ –	1/ 1	–/ –	11/ 16	69 %
6	abces	7/13	–/ 1	1/ 1	3/ 3	–/ –	1/ 1	2/ 3	–/ –	14/ 22	64 %
7	elmundo	4/ 5	2/ 3	8/ 8	–/ –	–/ –	–/ –	4/ 4	–/ –	18/ 20	90 %
8	lesechos	2/ 3	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	2/ 3	67 %
9	ledevoir	16/19	2/ 8	4/ 4	2/ 2	–/ –	1/ 2	3/ 3	–/ –	28/ 38	74 %
10	hvg.hu	2/ 2	–/ –	1/ 4	4/ 4	–/ –	1/ 2	2/ 2	–/ –	10/ 14	71 %
11	nemzet.hu	–/ –	1/ 6	–/ –	–/ –	–/ –	–/ –	–/ –	–/ –	1/ 6	17 %
12	adnkronos	–/ –	2/ 2	–/ –	1/ 2	–/ –	–/ –	2/ 3	–/ –	5/ 7	71 %
13	corriere	2/ 3	–/ 1	–/ –	–/ –	–/ –	–/ –	1/ 1	–/ –	3/ 5	60 %
		49/68	16/49	17/22	17/22	1/ 4	3/ 11	18/20	1/ 3	122/199	61 %
		72 %	33 %	77 %	77 %	25 %	27 %	90 %	33 %	61 %	

Table 1: Correct translations and total number of German anaphoric pronouns in a subset of the WMT 2009 test set.

in our case study is drawn from the German-English corpus used as a test set for the MT shared task at the EACL 2009 Workshop on Machine Translation [4]. The test set is composed of 111 newswire documents from various sources in German and English translations. In the selected subset of 13 documents (219 sentences) we identified all cases of pronominal anaphora that could be resolved in the text. One of the documents did not contain any such cases. For each anaphoric pronoun in the German source text, we manually checked whether or not it was translated into English in an appropriate way by our phrase-based SMT system submitted to the WMT 2010 shared task [5]. The system uses 6-gram language models, allowing it to consider a relatively large local context in translation, but it does not contain any specific components to process sentence-wide or cross-sentence context.

As can be seen in table 1, the MT system finds a suitable translation for anaphoric pronouns in about 61 % of the cases in this sample. The success of the MT system is strongly dependent on the type of pronoun: While it produces adequate output for around 90 % of the demonstrative pronouns (*dieser*, *dieses*, etc.) and about 3 out of 4 masculine or neuter singular pronouns or plural pronouns, only a third of the feminine pronouns are translated correctly. For pronouns of polite address and reflexive pronouns, the system largely fails.

The reasons for these discrepancies can most likely be found in the differences of the pronominal systems of the source and the target languages. There is no one-to-one correspondence between the German and the English singular pronouns. Moreover, some German pronouns are highly ambiguous. Thus, the pronoun *sie* can be the form of the feminine singular, of the plural of any gender or, when capitalised, of the polite form of address, which has to be translated into an English second person *you*. The reflexive pronoun *sich* is used for all genders and both numbers in the third person; it frequently has no direct equivalent in the English sentence. In these ambiguous cases, the language model will try to disambiguate based on parts of the context that were seen during training. If the local context is truly ambiguous, the results of this disambiguation will be essentially random. Generally, the system will prefer the forms that were observed most frequently at training time. For instance, it will tend to translate *sie* as a plural pronoun even when it is a feminine singular in reality.

Even though translation mistakes due to wrong pronoun choice do not generally affect important content words, they can make the MT output hard to understand, as in the following example from document 3 of our sample:

> *Input:* Der Strafgerichtshof in Truro erfuhr, dass *er seine* Stieftochter Stephanie Randle regelmässig fesselte, als *sie* zwischen fünf und sieben Jahre als [recte: alt] war.
>
> *Reference translation:* Truro Crown Court heard *he* regularly tied up *his* step-daughter Stephanie Randle, when *she* was aged between five and seven.
>
> *MT output:* The Criminal Court in Truro was told *it* was *his* Stieftochter Stephanie Randle tied as *they* regularly between five and seven years.

The MT output for this sentence suffers from several deficiencies, and bad pronoun choice is clearly part of them.

To sum up, there is evidence that current phrase-based Statistical MT cannot handle pronoun choice adequately. Al-

though the present case study is limited to a single language pair and a single text genre, considering the models used in SMT, there is no reason to suppose that the situation should be very different in other cases. Stronger differences in pronoun systems and text with longer, more complex sentences are likely to exacerbate the difficulties, whereas the problem will be easier to solve when the languages are close and the sentences are simple and homogeneous with the training corpus.

The results of this case study indicate that better handling of pronominal anaphora may lead to observable improvements in translation quality. In the remainder of this paper we describe and discuss an attempt to address this challenge by integrating the output of an automatic coreference resolver into our SMT system by means of a word dependency model modelling links between pairs of potentially remote words in the input text.

3. Previous work

There is little literature about modelling cross-sentential phenomena in Machine Translation, and most of it is relatively old. Anaphora resolution was a topic of interest in the literature on Rule-based Machine Translation (RBMT) in the 1990s. While the analyses of linguistic phenomena made for RBMT remain valid for any kind of MT activity, the problems encountered in RBMT output are different from the typical problems of SMT output, and approaches taken in this field generally rely on the system architecture of a rule-based transfer system and are not directly applicable to SMT. This strand of research culminated with the publication of a special issue of the journal *Machine Translation* on "Anaphora Resolution in Machine Translation and Multilingual NLP" in 1999 [6]. After this date, publication activity ebbed away.

In the SMT literature, the problem of translating pronominal anaphora was only taken up very recently by Le Nagard and Koehn [7]. Translating English into French, they use a coreference resolution system to label English pronouns in the training and the test corpus with the French gender of their French antecedents. They report unchanged BLEU scores. Manual evaluation reveals that the number of correctly translated pronouns slightly decreases from 69 % to 68 % when applying their procedure, which the authors put down to the low performance of their coreference resolution system. To our knowledge, this is the only attempt to handle anaphora explicitly in SMT in current literature.

4. Integrating Coreference Links into SMT

4.1. Coreference Annotation

In general, the decision which pronoun to emit in the target language cannot be taken based on local information only. In many languages, pronouns show complex patterns of agreement, and selecting the correct word form requires dependencies on potentially remote words. German possessive pronouns, for instance, agree in gender and number with the possessor (determining the choice between *sein*, *ihr*, etc.) and in gender, number and case with the possessed object (with a paradigmatic choice between, e. g., *sein*, *seine*, *seines*, etc., if the possessor is masculine singular). While the possessed object occurs in the same noun phrase as the pronoun and agreement can, at least in simpler cases, be enforced by an *n*-gram language model, the possessor can occur anywhere in the text, even in a different sentence. Since a given input word can be translated with different words in the target language and the pronoun must agree with the word that was actually chosen, correct pronoun choice depends on a translation decision taken earlier by the Machine Translation system. Our model attempts to face this challenge by explicitly identifying anaphoric links in the SMT input and measuring in the output how well the translation of an anaphoric pronoun matches the translation of its antecedent.

We used the open-source coreference resolution system BART [8] to link pronouns to their antecedents in the text. The preliminary case study described in the preceding section was about German-English translation. In our practical experiments, we worked on the inverse translation direction, English-German, because we had ready access to an English coreference resolver.

The coreference resolution system we used was trained on the *ACE02-npaper* corpus and uses separate models for pronouns and non-pronouns in order to increase pronoun-resolution performance. For each resolvable pronoun, the system finds a link to exactly one direct antecedent noun phrase. In our system, we use word-to-word links from the referent pronouns to the syntactic heads of the antecedent noun phrases. The output of the coreference resolver is illustrated in the upper part of figure 2, where the markable noun phrases are enclosed in square brackets and their syntactic heads highlighted in bold face. Information about complete coreference chains, also output by the coreference resolution system, was not used in our experiments.

4.2. Managing Cross-Sentence Dependencies

Handling cross-sentence coreference links requires propagating information from the translation output of one sentence to the input of a following sentence. Whenever a sentence contains a mention which is referred to anaphorically in a later sentence, the words chosen to translate this mention must be extracted and fed into the decoding process when the referring sentence is translated. We implemented a driver module that feeds the decoder and parses its output to manage this task.

Figure 1 illustrates the workings of the decoder driver. The output of the coreference resolution system is represented as a directed graph of sentences with their dependencies (top right). At the sentence level, we only use anaphoric links; cataphoric links, which are much rarer, are disregarded. This restriction guarantees that the sentence dependency graph is acyclic. Each sentence can contain pronominal mentions that refer to a preceding sentence (outgo-

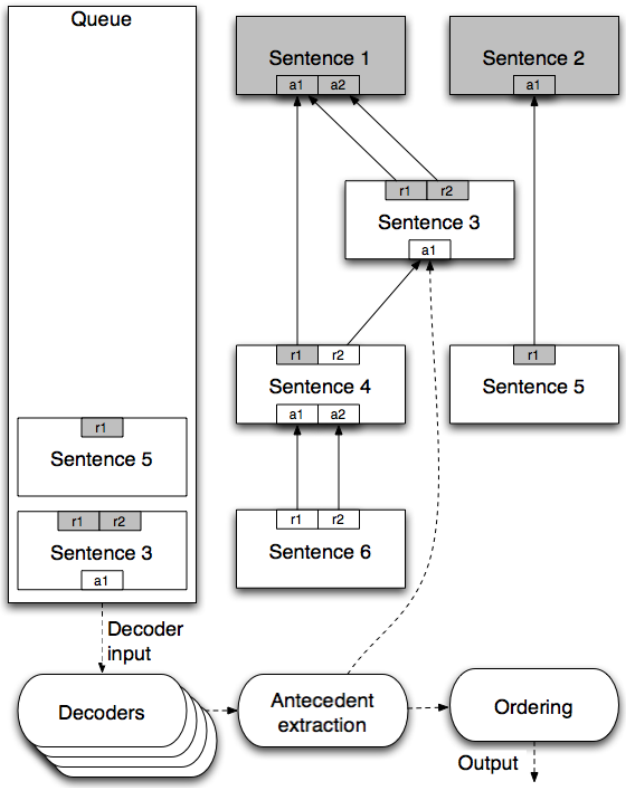

Figure 1: Managing cross-sentence dependencies for the decoder input.

ing dependencies, marked *r*) as well as antecedent mentions that are referred to later (incoming dependencies, marked *a*). The figure shows the state after translating sentences 1 and 2. Sentences that have no outgoing dependencies, such as sentences 1 and 2 in the example, and sentences whose outgoing dependencies have already been resolved, such as sentences 3 and 5, are put on a queue that feeds the decoder. After decoding, the translations of the antecedent mentions are recovered from the decoder output with the help of the phrase alignments produced by the decoder and the word alignments stored in the SMT phrase table. The decoder driver extracts the words aligned to what has been identified as the syntactic head of the antecedent mention and makes them available to the referring sentences by encoding them in the decoder input as described in the following section. Whenever all outgoing dependencies of a sentence are satisfied, the sentence is put on the queue.

The actual implementation is multi-threaded and feeds a number of parallel decoder processes. The decoder input queue is realised as a priority queue ordered by the number of incoming dependencies of the sentences in order to resolve as many dependencies as possible as early as possible and thus increase the throughput of the system. Since the sentences are not processed in order, a final ordering step restores the original document order.

4.3. Word-Dependency Model: Decoding

Coreference information was integrated into an SMT system based on the phrase-based Moses decoder [9] in the form of a new model which represents dependencies between pairs of target-language words produced by the MT system. The decoder driver encodes the links found by the coreference resolver in the input passed to the SMT decoder. Pronouns and their antecedents are marked as illustrated in the lower half of figure 2. Each token is annotated with a pair of elements. The first part numbers the antecedents to which there is a reference in the same sentence. The second part contains the number of the sentence-internal antecedent to which this word refers, or the word itself, if it occurred in a previous sentence. Each part can be empty, in which case it is filled with an asterisk.

Instead of using the word forms for our word dependency model, we map the antecedent words to a tag representing their gender and number; thus, in the example, the word *hospital* in the first sentence, which is translated by the system into the neuter singular word *Krankenhaus* (not shown), gets mapped to the tag `neut_sg` in the input for sentence 2. Gender and number of German words were annotated using the RFTagger [**?**]. The representation of the referent words, by contrast, is fully lexicalised.

The word dependency module is integrated as an additional feature function in a standard log-linear SMT model [10]. It keeps track of pairs of source words $(s_{\text{ant}}, s_{\text{ref}})$ participating as antecedent and referent in a coreference link. Usually, the antecedent s_{ant} will be processed first; however, it is also possible for the referent s_{ref} to be encountered first, either because of a cataphoric link in the source sentence or, more likely, because of word reordering during decoding. When the second element in an antecedent-referent pair is translated, the word dependency module adds a score of the following form:

$$p(T_{\text{ref}}|T_{\text{ant}}) = \max_{(t_{\text{ref}},t_{\text{ant}})\in T_{\text{ref}}\times T_{\text{ant}}} p(t_{\text{ref}}|t_{\text{ant}}), \quad (1)$$

where T_{ref} is the set of target words aligned to the source word s_{ref} and T_{ant} is the set of target words aligned to the source word s_{ant} in the decoder output. Word alignments between decoder input and decoder output are reconstructed based on the phrase-internal word alignments that led to the extraction of the phrases during SMT system training.

Coreference links across sentence boundaries are handled by a special module that reads the decoder output and extracts the required information about antecedents occurring in previous sentences, encoding it in the input of the sentence containing the reference as described above. In this case, the antecedent is not marked in the decoder input, but silently extracted from the output, and the referent token is decorated directly with the extracted word form. Cataphoric links across sentence boundaries are not handled by the model.

During the search process in the SMT decoder, a search path can be abandoned when the decoder can prove that there

[The same **hospital**]$_1$ had had to contend with a similar infection early this year. [**It**]$_{2;\ \text{ant:1}}$ had discharged a patient admitted after a serious traffic accident. Shortly afterward, [**it**]$_{3;\ \text{ant:2}}$ had to re-admit the patient because of an MRSA infection, and [**doctors**]$_4$ have been unable to perform surgery that would be vital to full recovery because [**they**]$_{5;\ \text{ant:4}}$ have been unable to get rid of the staph.

```
The same hospital had had to contend with a similar infection early this year .
It|*->neut_sg had discharged a patient admitted after a serious traffic accident .
Shortly afterward , it|*->neut_sg had to re-admit the patient because of an MRSA
infection , and doctors|1-* have been unable to perform surgery that would be
vital to full recovery because they|*-1 have been unable to get rid of the staph .
```

Figure 2: Coreference link annotation and decoder input

is another search path that is superior under every possible continuation of the search. This is called hypothesis recombination [11]. Since our model introduces dependencies that can span large parts of the sentence, care must be taken not to recombine hypotheses that could be ranked differently after including the word dependency scores. We therefore extend the decoder search state to include, on the one hand, the set of antecedents already processed and, on the other hand, the set of referents encountered for which no antecedent has been seen yet. In either case, the translation chosen by the decoder is stored along with the item. Hypotheses can only be recombined if both of these sets match.

4.4. Word-Dependency Model: Training

The probability distribution $p(t_{\text{ref}}|t_{\text{ant}})$ in equation 1 is estimated as a bigram language model. Training examples are extracted from a parallel corpus in a way similar to the application of the model: The source language part of a word-aligned parallel corpus is annotated for coreference with the BART software, then the antecedent and referent words are projected into the target language using the word alignments and the corresponding pairs of target-language antecedent and referent words are used as training examples. A plausible alternative would be to train the model directly on coreference pairs extracted in the target language.

Our model was trained on the news-commentary10 corpus provided as training data for the WMT shared tasks. The estimated probabilities were smoothed using the Witten-Bell method [12]. This smoothing method does not make prior assumptions about the distribution of n-grams in a text. It is therefore more suited for estimating the probabilities of events not drawn directly as n-grams from a text than the Improved Kneser-Ney method we used for smoothing our other n-gram models.

5. Evaluating Pronoun Translation

Since our model addresses a specific problem of the MT system, evaluation with a general-purpose score such as BLEU may not be fully adequate. Besides measuring overall translation quality, which is what general-purpose measures purport to do, we also want to know details about the impact of the new model on pronoun translation. We therefore propose a method to measure precision and recall of pronoun translations more directly.

We use a test corpus with one reference translation, for which we construct word alignments by concatenating it with additional parallel training data, running the GIZA++ word aligner [13] and symmetrising the alignments as is usually done for SMT system training. We also produce word alignments between the source text and the candidate translation by considering the phrase-internal word alignments stored in the phrase table. The basic idea is to count the number of pronouns translated correctly. Doing so would require a $1:1$ mapping from pronouns to their translations. However, word alignments can link a word to zero, one or more words, so we suggest using a measure based on precision and recall instead.

For every pronoun occurring in the source text, we obtain the set of aligned target words in the reference and the candidate translation, R and C, respectively. Inspired by the BLEU score [14], we define the clipped count of a particular candidate word w as the number of times it occurs in the candidate set, limited by the number of times it occurs in the reference set:

$$c_{\text{clip}}(w) = \min\left(c_C(w), c_R(w)\right) \tag{2}$$

We then consider the match count to be the sum of the clipped counts over all words in the candidate translation aligned to pronouns in the source text, which allows us to define precision and recall in the usual way:

$$\text{Precision} = \frac{\sum_{w \in C} c_{\text{clip}}(w)}{|C|}; \text{Recall} = \frac{\sum_{w \in C} c_{\text{clip}}(w)}{|R|} \tag{3}$$

This measure can be applied both to obtain a comprehensive score for a particular system on a test set or to compute detailed scores per pronoun type to gain further insights into the workings of the model.

For testing the significance of recall differences, we used a paired t-test. Pairing was done at the level of the set R, the individual target words aligned to pronouns in the reference translation. This method is not applicable to precision, as the

	newstest 2008	newstest 2009
Baseline		
Precision	33.3 %	42.8 %
Recall	30.2 %	38.8 %
F1	31.7 %	40.7 %
Word-dependency model		
Precision	33.8 %	43.0 %
Recall	31.6 %	39.9 %
F1	32.6 %	41.4 %

Table 2: Pronoun translation precision and recall

sets C cannot be paired among different candidate translation.

6. Experimental results

The baseline system for our experiments was built for the English-German task of the ACL 2010 Workshop on Machine Translation. It is a phrase-based SMT system based on the Moses decoder with phrase tables trained on the Europarl version 5 and news-commentary10 parallel corpora and a 6-gram language model trained on the monolingual News corpus provided by the workshop organisers with the IRSTLM language modelling toolkit [15].

The feature weights were optimised by running Minimum Error-Rate Training (MERT; [**?**]) over the news-test2008 development set for the baseline system. In order to minimise the influence of feature weight selection on the outcome of the experiments, we did not rerun MERT when adding the word dependency model. Instead, we reused the baseline feature weights and conducted a grid search over a set of possible values for the weight of the word dependency model, selecting the setup that yielded best pronoun translation F-score on news-test2008. The optimal weight was found to be 0.05 with the other 14 weights (7 distortion weights, 1 language model, 5 translation model weights and word penalty as in a baseline Moses setup) normalised to sum 1.

English-German is a relatively difficult language pair for SMT because of pervasive differences in word order and very productive compounding processes in German. Our baseline system achieves a BLEU score of 13.66 % on the newstest-2009 test set. The best system submitted to WMT 2009 scored 14.8 % on the same test set. Handling pronouns with a word dependency model had no significant effect on the BLEU scores, which varied between 13.6 % and 13.7 % in all our experiments.

The pronoun-specific evaluation (table 2) clearly shows that the SMT system is very bad at translating pronouns in general. Indeed, most of the pronouns are not translated correctly. For both test sets, adding the word dependency model results in a tiny improvement in precision and a small improvement in recall, which is however highly significant ($p < .0005$ in a one-tailed t-test for both test sets).

A closer look at the performance of the system on individual pronouns reveals that by far the largest part of the improvement stems from the pronoun *it*, which is translated significantly better by the enhanced system than by the baseline. Recall for this pronoun improves from 21.02 % to 27.08 % for the news-test2008 corpus ($p < .0001$, two-tailed t-test) and from 21.80 % to 25.06 % for the newstest2009 corpus ($p < .005$). The only other item which benefits from a significant improvement at a confidence level of 95 % is, surprisingly enough, the first-person pronoun *I* in the newstest2009 corpus (from 60.40 % to 62.40 %, $p < .05$). In the news-test2008 corpus, the word dependency model has no effect whatever on the word *I*, so it seems likely that this improvement is accidental.

By contrast, the improvement we obtain for the pronoun *it*, albeit slight, is encouraging. While most other English pronouns such as *he*, *she*, *they* etc. are fairly unambiguous when translated into German and the ambiguity the MT system is faced with will mostly concern case marking or the difficult question whether or not a pronoun is to be translated as a pronoun at all, translating *it* requires the system to determine the grammatical gender of the German antecedent in order to choose the right pronoun. Similar problems occur in the opposite translation direction and in other language pairs, e. g. when translating the highly ambiguous German pronoun *sie* into English, or when translating between two languages that have different systems of grammatical gender.

7. Discussion and Conclusion

In the present paper, we have presented a novel approach to the problem of dependencies spanning sentence boundaries. This problem has long been neglected in SMT research, yet it is unavoidable if an MT system is supposed to translate a text as a text rather than as a bag of unconnected sentences. We introduced a word dependency model that handles long-range dependencies between pairs of words and presented a framework that makes it possible to handle dependencies from a word in one sentence to another word in one of the preceding sentences. Applying this model to the pervasive problem of pronoun choice led to a small, but statistically significant improvement in the recall of pronoun translations with no noticeable negative effect on other aspects of translation quality. This is in contrast with previous research [7], which did not result in any measurable improvement at all. We are therefore confident that our model is a first step in the right direction.

We suspect that the low impact of our model on overall translation quality is partly due to the generally low quality of English-German SMT. As a next step, we therefore plan to apply our model to easier language pairs such as English-French or German-English. The method described in this paper can easily be applied to other language pairs, provided that there is a working anaphora resolution system. Another

problem is the sheer level of noise present in the system, especially in the word alignments and in the output of the coreference resolution system. We intend to quantify the latter problem by running similar experiments with manually annotated coreference links. Finally, we believe that the word dependency model proposed in this paper can also be used for other types of dependencies not linked to anaphora resolution, such as agreement of verb forms or adjectives.

We are convinced that extra-sentential dependencies cannot be avoided in SMT research much longer. What we have presented is a step towards handling this problem, and we hope that more work on this topic will ensue.

8. Acknowledgements

We are indebted to Olga Uryupina for providing us with a fully trained coreference resolution model for English and helping us to set up the BART coreference resolver. This work was partially funded by the 7th Framework Programme of the European Commission through the contract T4ME, grant agreement no.: 249119.

9. References

[1] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, *et al.*, "A statistical approach to Machine Translation," *Computational linguistics*, vol. 16, pp. 79–85, 1990.

[2] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of Statistical Machine Translation," *Computational linguistics*, vol. 19, pp. 263–311, 1993.

[3] M. Strube, "Anaphora and coreference resolution, Statistical," in *Encyclopedia of language and linguistics*. Elsevier, 2006, pp. 216–222.

[4] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder, "Findings of the 2009 Workshop on Statistical Machine Translation," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece: Association for Computational Linguistics, March 2009, pp. 1–28.

[5] C. Hardmeier, A. Bisazza, and M. Federico, "FBK at WMT 2010: Word lattices for morphological reduction and chunk-based reordering," in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*. Uppsala, Sweden: Association for Computational Linguistics, 2010.

[6] R. Mitkov, "Introduction: Special issue on anaphora resolution in Machine Translation and Multilingual NLP," *Machine translation*, vol. 14, pp. 159–161, 1999.

[7] 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*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 252–261.

[8] S. Broscheit, M. Poesio, S. P. Ponzetto, K. Joseba Rodriguez, L. Romano, O. Uryupina, Y. Versley, and R. Zanoli, "BART: A multilingual anaphora resolution system," in *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010),* Uppsala, Sweden, 15–16 July 2010, 2010.

[9] P. Koehn, H. Hoang, A. Birch, *et al.*, "Moses: open source toolkit for Statistical Machine Translation," in *Annual meeting of the Association for Computational Linguistics: Demonstration session*, Prague, 2007, pp. 177–180.

[10] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for Statistical Machine Translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, Philadelphia, 2002, pp. 295–302.

[11] F. J. Och, N. Ueffing, and H. Ney, "An efficient A* search algorithm for Statistical Machine Translation," in *Proceedings of the Data-Driven Machine Translation Workshop, 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, Toulouse, 2001, pp. 55–62.

[12] I. H. Witten and T. C. Bell, "The zero-frequency problem: Estimating probabilities of novel events in adaptive text compression," *IEEE Transactions on Information Theory*, vol. 37, pp. 1085–1094, 1991.

[13] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*, vol. 29, pp. 19–51, 2003.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of Machine Translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Philadelphia: ACL, 2002, pp. 311–318.

[15] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an open source toolkit for handling large scale language models," in *Interspeech 2008*. ISCA, 2008, pp. 1618–1621.