WDS'11 Proceedings of Contributed Papers, Part I, 155–160, 2011. ISBN 978-80-7378-184-2 © MATFYZPRESS

Utilization of Anaphora in Machine Translation

M. Novák
Charles University in Prague, Faculty of Mathematics and Physics,
Malostranské nám. 25, Praha 1, CZ-118 00, Czech Republic.

Abstract. Majority of present machine translation systems do not address the retaining of text coherency, they translate just isolated sentences. On the other hand, the authors of anaphora resolvers rarely integrate these tools into more complex scenarios, e.g. the task of machine translation. We propose the ways how machine translation systems can utilize the knowledge of anaphoric relations both in the source as well as in the target language in order to improve the quality of translation. Specifically, we present how to incorporate anaphora resolution into the process of English to Czech translation using the TectoMT system.

Introduction

Machine translation (MT) is probably one of the most established tasks in the field of natural language processing. Therefore, it is surprising that nowadays so little attention is received by research on retaining the text coherency. On the contrary, current best systems based on statistical machine translations (SMT) make some strong independence assumptions.

Although very popular phrase-based SMT systems do not usually assume independence on segments smaller than a sentence, they make use of n-gram statistics,[1] which can capture only adjacent words and phrases. These systems face problems with handling long distance dependencies.

This shortcoming can be minimized by using SMT systems that include linguistic analysis up to the layer of deep syntax (e.g. TectoMT [Žabokrtský et al., 2008]). On this layer, a sentence is usually represented as a tree, keeping track of dependencies between the words, no matter how far from each other they appeared on the surface.

However, even for this deeper SMT approach, some problems still persist. The largest issue lies in that SMT systems translate individual sentences without taking the context of previous sentences into account. Another shortcoming, for example in TectoMT, is that even though it well describes dependency relations including the long distance ones, it insufficiently covers other types of relations, e.g. discourse and anaphoric relations.

Error analysis of TectoMT translation

In order to justify the claims on SMT shortcomings above, we carried out a survey on the English sentences translated by TectoMT system into Czech to find out how these problems reflect in real data. We made use of the data from the English-Czech test set for Shared Translation Task organized with EMNLP 2011 Sixth Workshop on Statistical Machine Translation.[2]

One of the problems that reduces the coherency of the translated text is the translation of personal pronouns. SMT systems usually do not investigate to what a pronoun refers. It results in an inconsistency between the translated pronoun and the translated referent.

To show the shortcomings of SMT in pronoun translation we conducted the following analysis (illustrated in Table). We randomly selected 100 English sentences that contained a pronoun "it", comprising 119 occurrences of this pronoun in total. We observed that more than half of them corresponds to either references to a bigger segment of the text (a clause, sentence or paragraph[3]) or pleonastic usages.[4] In the former case after their translation into the Czech variant in neutral almost never an error is incurred and in the latter case it is rather a matter of syntax how the pronoun's governing phrase looks like in the target language. Thus, we were not interested in these occurrences.

The rest of the occurrences referred to some entity mentioned in the previous text and we examined how the TectoMT system succeeded in their translation. Even though the English pronoun "it" can be translated into one of the three Czech pronouns, which differ in a gender, TectoMT outputs it always in neutral. We observed that the number of erroneously translated pronouns "it" account for 26% of

[1] With n usually smaller than 5.
[2] http://www.statmt.org/wmt11/test.tgz
[3] Boundaries of the segment the pronoun refers to does not even have to be precisely defined.
[4] We categorized them according to rules proposed in [Li et al., 2009].

	Deep	**Surface**	**Total**
	Erroneous translation		
Fem	9	18	27
Masc	4	13	17
Total	13	31	44
	Correct translation		
Neut	1	19	20
	Entity non-referring		
Pleo	—		31
Segm	—		24
Total	—		55

Table 1. The results of analysis of translating 119 pronouns "it". The section *Erroneous translation* contains pronouns incorrectly translated to neutral gender. The label Deep denotes those, which by chance did not produce a mistake on the surface but they were erroneous on the deep syntactic layer. The label Surface denotes those, which produced error also on the surface. The section *Correct translation* contains correctly translated pronouns. The label Deep denotes those, which were correctly translated on the deep syntactic layer, however, during the synthesis an error was incurred. The label Surface denotes those, which were correct also on the surface. The section *Entity non-referring* denotes pleonastic and segment-referring pronouns.

(a) **The leader of the PPC** ... present **her** conditions ... ***The president*** **of the popular Catalans** *went* to
Ref: Předsedkyně/fem ... se dostavila/fem
Tst: Prezident/masc ... šel/masc

(b) It is a combination of **location detection via GPS** (used by regular **car** ***navigation***) ...
Ref: navigace
Tst: plavba

(c) ... ***bunch*** of ripe **bananas** ...
Ref: svazek
Tst: parta

Figure 1. Illustrating examples from the English to Czech translation by TectoMT, where text coherency was not retained. Labels "Ref" and "Tst" denote a correct reference translation and a translation output by TectoMT, respectively.

all occurrences and together with errors that by chance did not appear on the surface[5] they form 37% (more than 2/3 of those referring to some entity).

Regarding the wrong choice of a gender we also came across the example in Figure 1a. The phrase "president ... went to" was mistakenly translated into masculine gender. If the system took into account that expressions in bold denote the same object and one of the expressions was a possessive pronoun "her", it would correctly output the subject with verb in feminine gender.

In the output of TectoMT system, we also noticed the incorrect choice of the translation for some word in the source text, caused by the lack of knowledge about the previous context. For instance, in the example in Figure 1b despite the evidence in the context that the sentences describe a car navigation system and GPS technology, the word "navigation" was erroneously translated into "plavba" ("cruise"). Similarly, in the example in Figure 1c the system failed in the translation of the word "bunch" into "parta" ("group" in the meaning of "group of people").

From the analysis and some particular examples above we can see that TectoMT system suffers from not taking a previous context into account and from insufficient handling of other than dependency relations, e.g. the anaphoric relations.

[5]For instance, a verb in the present tense has the same form no matter in what gender the subject is. However, it no longer holds for verbs in the past tense.

Anaphora as a means of coherence

In this paper we would like to address the mentioned deficiencies and suggest some solutions to avoid them. We believe that the awareness of anaphoric relations, i.e. relations between entities that appear in a text, in the process of machine translation could help in minimizing the errors.

Let us return to the examples found in the data. In the example in Figure 1a all the expressions in bold refer to the same entity – the leader of a Catalan popular party. They represent a special type of anaphora called coreference. In the example in Figure 1b the "location detection via GPS" is a function of a "car navigation", thus these expressions form a function–object bridging anaphora. Similarly, in the last example in Figure 1c "banana" is a part of the whole "bunch", what is in the theory of anaphora[6] denoted as a part–whole bridging anaphora. From these examples we see that if we had a tool to reveal such relations, we could use them in the process of MT to retain the coherency of the translated text, thus improving the quality of the translation.

Related work

A lot of research has been carried out on anaphora resolution (AR), especially on the coreference resolution (CR).[7] Many of hundreds of works on AR published so far declared MT as one of their main motivation. In light of that it seems peculiar that almost none of them conducted any experiments on integration of AR into the MT system.

The lack of interest in utilizing anaphora knowledge in MT was not present all the time. During 1990s some authors of rule-based MT systems attempted to handle inter-sentential relations. In 1999 this effort culminated with a special issue of Machine Translation journal concerning anaphora resolution in MT [Mitkov, 1999]. An example of such rule-based system using AR can be the work of Peral and Rodríguez [2002], who implemented translation from English to Spanish and vice versa using transfer via interlingua.

After the rise of SMT approach the research on this topic paused for 10 years. Just recently two new works have emerged, both applying CR on translation of personal pronouns (especially "it" and "they") from English into French and German, respectively. Le Nagard and Koehn [2010] tagged pronoun "it" with gender information by replacing it with one of the following surface forms: "it-neutral", "it-feminine" and "it-masculine". The corresponding form of corefering pronoun is determined by the gender of the antecedent's French translation. On the other hand Hardmeier and Federico [2010] introduced a robust system of translating the referred sentences into German in advance and integrated a so-called word dependency module comprising information on coreference links as an additional feature into a log-linear SMT model. Whereas results of the former suffered from usage of low-performance rule-based systems for CR, in the latter work they succeeded in increasing the quality of personal pronoun translation.[8]

Our suggestions

In the remaining part of this paper we present our suggestions on how to incorporate anaphora knowledge into the process of SMT, which might help in retaining coherency in the output of the translation. Our proposals are aimed to be integrated into TectoMT system, especially into the English to Czech translation via deep syntactic (tectogrammatical) transfer.

We opted for this language pair and direction because the tools and models neccessary for English to Czech translation are the most available for us. The majority of our suggestions requires the anaphora resolver for the source language. English seems to be the best choice, since a huge number of CR system are implemented. In addition, both languages are for us easy to understand, which facilitates revealing of errors.

The main reason for translation via a tectogrammatical layer, i.e. using the TectoMT framework, lies in a design of an available CR system for Czech. To our knowledge the only existing system presented by Nguy et al. [2009] requires several features, which are not present on linguistic layers lower than the deep syntactic layer.

Note, however, that the following suggestions were designed to be applied for any language pair. In addition, following ideas can be employed in a slightly modified way even for phrase-based SMT (for instance, in the system Moses).

[6] The typology we use is based on the theory described by Nědolužko [2009].
[7] The summary of statistical methods on CR can be found for instance in [Ng, 2010].
[8] Nonetheless, they did not achieve improvement in BLEU [Hardmeier and Federico, 2010].

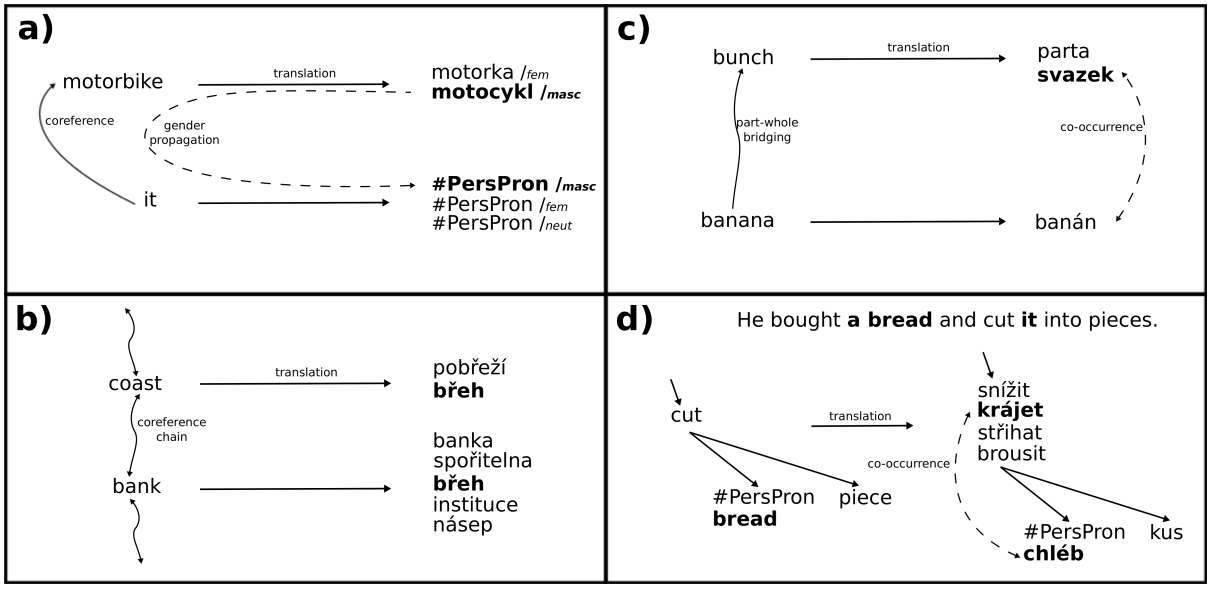

Figure 2. Suggestions of anaphora utilizing in MT on the side of the source language. The lemma #PersPron is an abstraction of all personal and possesive pronouns.

Anaphora resolver on the source side

AR resolver can be utilized on the side of the source language in several ways. One of them was already proposed in the works of Le Nagard and Koehn [2010] and Hardmeier and Federico [2010], i.e. helping to solve the issue of personal pronoun translation. To facilitate the correct translation of a personal pronoun (e.g. "it" in Figure 2a), the system has to be provided with the gender[9] of the closest antecedent's translation (e.g. masculine, if "motorbike" is translated to "motocykl"). This requires to translate referred sentences before the referring ones and to keep track of a gender of an antecedent's translation (propagated gender). Then we can determine the gender of the translated pronoun to agree with the propagated one or allow some value of uncertainty by introducing an additional feature, which describes this agreement, into a log-linear model.

Whereas in the utilization of anaphora mentioned above the system needs information just from the last coreferential expression preceding the pronoun, in the following idea we suppose we can obtain better results, if the available coreference chain is longer. We can use the knowledge of a whole coreference chain, i.e. all expressions referring to the same entity, to pick translation equivalents more confidently. When searching for the Czech translation of an English word e, instead of selecting the Czech word c_i that maximizes the probability $p(c_i|e)$, we choose the word c_j that maximizes the probability of translating e to c_j in the context of whole entity E that e belongs to. We calculate this probability as a weighted sum of probabilities that c_j is a correct translation of m over all expressions m that belongs to E. Written in a formula:

$$p_E(c_j|e) = \sum_{m \in E} \lambda_{e,m} p(c_j|m),$$

where the dependence of weights $\lambda_{e,m}$ also on the current translated word e allows for instance to prefer the translation probability $p(c_j|e)$ of the word e. Considering the example illustrated in Figure 2b, if the SMT system takes the whole coreference chain into account, it can favor "břeh" ("coast") over "banka" ("bank" as an institution) as a translation of English word "bank".

Similarly, we can use the knowledge of bridging anaphora in the same manner, as it is depicted in Figure 2c. Given the whole-part relation between the tectogrammatical nodes "bunch" and "banana", we can copy this relation into the target tree and infer the translations of nodes with respect to the constraint set by the relation (the system picks "svazek" rather than "parta" as a translation of "bunch"). More generally, it does not have to be strictly defined what kind of relation joins the words, it should be enough just to know whether such a relation between them exists. For setting these relations we can employ some of the association measures proposed for example in [Pecina, 2008].

The last proposal, how coreference knowledge on the source side could improve the SMT, is the help

[9] Sometimes the grammatical number can be required for correct translation as well.

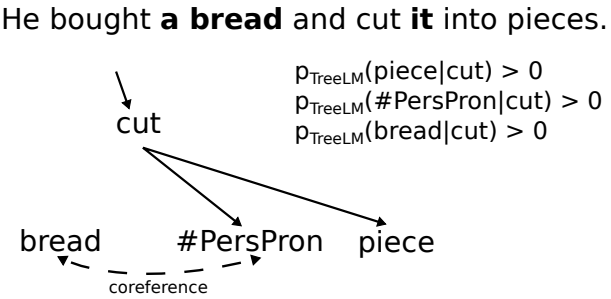

Figure 3. A suggestion of enrichment of target-language tree model, if taking the coreference relations into account. For clarity, we used an English example. In English – Czech translation that we focus on, the model would be nonetheless built from Czech data.

to decide the translation of a translated node's parent or children in a tectogrammatical tree. Considering again an example sentence "He bought a bread and cut it into pieces" (Figure 2d), the translated pronoun "it" alone does not give us sufficient information to predict the appropriate translation of the verb "cut". However, if we replace the node "#PersPron" of the pronoun expression by the node of the whole entity ("bread", "#PersPron"), its translation ("chléb", "#PersPron") then gives preference to "krájet" as the translation of the verb.

Anaphora resolver on the target side

While it is more obvious how to integrate anaphora resolution on the side of the source language, it is probably not so clear how to exploit the knowledge served by anaphora resolver on the side of the target language.

We suggest to make use of anaphora knowledge for in target-language tree model (TreeLM) as described in [Mareček et al., 2010]. TreeLM specifies the probabilities of nodes' word forms[10] given the word forms of their parents. Let us assume that the data, which these probabilities are estimated from, are enriched with coreference relations. If we enable replacement of individual personal pronoun expressions with the expressions they are coreferential with, we can build a tree model not only from those couples that co-occurred within the parent – child dependency relation, but it allows also for inclusion of the couples, which can potentially appear.

For instance, in Figure 3 the data for TreeLM consist solely of one sentence, in which the nodes "bread" and "#PersPron" are coreferential. Then, given that the parent is a node "cut", TreeLM can predict non-zero probabilities not only for the node "#PersPron" but also for the node "bread".

Such an enrichment of the TreeLM can both give rise to new dependency relations that did not appear in the corpus and lead to more reliable estimates of probabilities assigned to dependency relations.

Conclusion

In this paper we gave some proposals, how the knowledge of anaphora relations could help to improve the quality of MT. While our contribution so far lies in the stage of suggestions, we plan to continue this project and test the methods on English – Czech translation. Our proposals should be nevertheless appliable to all language pairs and some of the proposed ideas can be easily adapted to other SMT strategies than the deep-syntactic transfer we assumed here.

Acknowledgments. This work was supported by the grants GAUK 4226/2011 and Czech Science Foundation 201/09/H057. We thank two anonymous reviewers for their useful comments.

References

Hardmeier, C. and Federico, M., Modelling Pronominal Anaphora in Statistical Machine Translation, in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, edited by M. Federico, I. Lane, M. Paul, and F. Yvon, pp. 283–289, 2010.

Le Nagard, R. and Koehn, P., Aiding Pronoun Translation with Co-Reference Resolution, in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pp. 252–261, Association for Computational Linguistics, Uppsala, Sweden, 2010.

[10] In fact, rather than of a word form it specifies the probabilities of two attributes, the word form can be decomposed to: the lemma and formeme, which captures the surface morphosyntactic form of the node.

Li, Y., Musílek, P., Reformat, M., and Wyard-Scott, L., Identification of Pleonastic It Using the Web, *J. Artif. Intell. Res. (JAIR)*, *34*, 339–389, 2009.

Mareček, D., Popel, M., and Žabokrtský, Z., Maximum entropy translation model in dependency-based MT framework, in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp. 201–201, Association for Computational Linguistics, Uppsala, Sweden, 2010.

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

Ng, V., Supervised Noun Phrase Coreference Research: The First Fifteen Years, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1396–1411, Association for Computational Linguistics, Uppsala, Sweden, 2010.

Nguy, G. L., Novák, V., and Žabokrtský, Z., Comparison of classification and ranking approaches to pronominal anaphora resolution in Czech, in *Proceedings of the SIGDIAL 2009 Conference*, pp. 276–285, Association for Computational Linguistics, London, UK, 2009.

Nědolužko, A., *Zpracování rozšířené textové koreference a asociační anafory na tektogramatické rovině v Pražském závislostním korpusu*, Ph.D. thesis, MFF UK, Praha, Czech Republic, in Czech, 2009.

Pecina, P., *Lexical Association Measures: Collocation Extraction*, Ph.D. thesis, Charles University in Prague, Prague, Czech Republic, 2008.

Peral, J. and Rodríguez, A. F., Pronominal Anaphora Generation in an English-Spanish MT Approach, in *CICLing*, pp. 187–196, 2002.

Žabokrtský, Z., Ptáček, J., and Pajas, P., TectoMT: Highly modular MT system with tectogrammatics used as transfer layer, in *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 167–170, 2008.