Two Case Studies on Translating Pronouns in a Deep Syntax Framework

Michal Novák, Zdeněk Žabokrtský and Anna Nedoluzhko
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800
{mnovak,zabokrtsky,nedoluzko}@ufal.mff.cuni.cz

Abstract

We focus on improving the translation of the English pronoun *it* and English reflexive pronouns in an English-Czech syntax-based machine translation framework. Our evaluation both from intrinsic and extrinsic perspective shows that adding specialized syntactic and coreference-related features leads to an improvement in translation quality.

1 Introduction

Machine Translation (MT) is an extremely broad task and can be decomposed along various directions. One of them lies in using specialized translation models (TMs) for certain types of language expressions. For instance, different types of named entities often receive specialized treatment in real translation systems. This paper deals with introducing specialized TMs for two types of pronouns: the pronoun *it* and reflexive pronouns. The models are integrated into an English-Czech syntax-based MT framework.

Several works have previously focused on translating pronouns. The linguistic study of Morin (2009) investigated the translation of pronouns, proper names and kinship terms from Indonesian into English. Onderková (2010) has conducted a corpus-based research on possesive pronouns in Czech and English, focusing especially on their use with parts of the human body.

From the perspective of MT, translating personal pronouns from English to morphologically richer languages, such as French (Le Nagard and Koehn, 2010), German (Hardmeier and Federico, 2010) and Czech (Guillou, 2012) has recently aroused higher interest. In these languages, one usually has to ensure agreement in gender and number between the pronoun and its direct antecedent, which requires a coreference resolver to be involved.

In this work, we make use of the English-to-Czech translation implemented within the TectoMT system (Žabokrtský et al., 2008). In contrast to the phrase-based approach (Koehn et al., 2003), TectoMT performs a tree-to-tree machine translation. An input English sentence is first analyzed into its deep-syntactic representation, which is subsequently transferred into Czech. The pipeline ends with generating a surface form of the Czech translation from its deep representation.

The deep syntactic representation of a sentence in TectoMT follows the Prague tectogrammatics theory (Sgall, 1967; Sgall et al., 1986). It is a dependency tree whose nodes correspond to content words. Personal pronouns missing on the surface are reconstructed in special nodes. All nodes are assigned semantic roles and coreference relations are annotated.

Originally, translation of both *it* and reflexive pronouns was treated by rules in TectoMT. The English deep representation of *it* was translated as *to* and a simple heuristics determined if it is being expressed on the surface. Similarly, reflexives were always translated as *se*. This paper evaluates the translation quality reached using specialized classifiers for these pronouns. Unlike the related work on pronouns in MT, we focus on improving the lexical choice, not tuning other components that affect generating a particular surface form (e.g. coreference resolution).

2 Linguistic analysis

We started with an analysis of how the pronouns under investigation are translated[1] in two Czech-English parallel treebanks – Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2011, PCEDT) and CzEng 1.0 (Bojar et al., 2012).

[1]Note that besides the means mentioned below, there are other ways of translating these pronouns. However, in most cases they can be replaced by one of the variants listed with no harm to the quality of the Czech output.

International Joint Conference on Natural Language Processing, pages 1037–1041,
Nagoya, Japan, 14-18 October 2013.

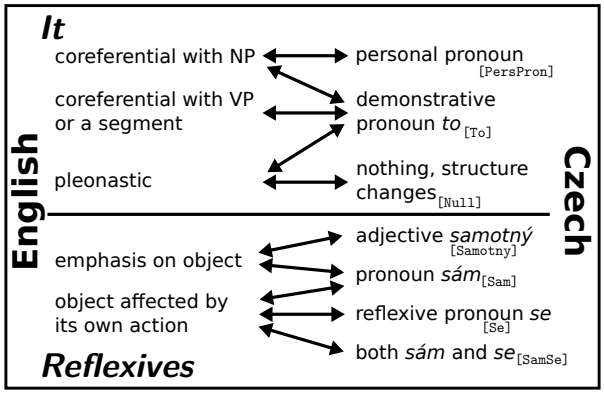

Figure 1: The mapping of the types of English *it* (top) and reflexive pronouns (bottom) to their Czech counterparts.

2.1 Translating *it* from English to Czech

In English, three coarse-grained types of *it* are traditionally distinguished: referential *it* pointing to a noun phrase in the preceding or following context, anaphoric *it* referring to a verbal phrase or a larger discourse segment, and non-referential pleonastic *it*, whose presence is imposed only by the syntactic rules of English.

There are three prevailing ways of translating *it* into Czech, also three different ways prevail. Personal pronouns or zero forms[2], whose gender and number are determined by their antecedent, are the most frequent variant (referred to as the `PersPron` class in the following). Another way is using the Czech demonstrative pronoun *to*, which is a neuter singular form of the pronoun *ten* (`To` class). The third option results in fact no lexical counterpart in the Czech translation, the English and Czech sentences thus having a different syntactic structure (`Null` class).

The mapping between English and Czech types is shown in Figure 1. The `To` class is particularly overloaded. Even if a given occurrence of *it* corefers with a noun phrase, translating it to *to* does not require identifying the antecedent since the gender and number of *to* are always fixed (see Example 1).

(1) Some investors say Friday's sell-off was a good thing. "*It* was a healthy cleansing," says Michael Holland.

Někteří investoři říkají, že páteční výprodej byla dobrá věc. "Byla *to* zdravá očista," říká Michael Holland.

[2]Czech is a pro-drop language.

2.2 Translating reflexive pronouns from English to Czech

According to the *Longman Dictionary of Contemporary English*,[3] reflexive pronouns are typically used in two scenarios: to show that the object is affected by its own action and to emphasize that the utterance relates to one particular thing, person etc. (see Example 2).

(2) The Gambia's President *himself* participated in the hunt last year.

The most usual Czech counterparts of English reflexives comprise the Czech reflexive pronoun *se* (`Se` class), the adjective *samotný* (`Samotny` class) and the pronoun *sám* (`Sam` class), all in various morphological forms. Moreover, *sám* often appears with *se* to emphasize that the action affecting the object is performed by the object itself (`SamSe` class). Figure 1 illustrates the correspondence between English usages and Czech expressions.

3 Data

To train and intrinsically evaluate TMs for *it* and English reflexives, we have extracted data from the entire PCEDT and 11 sections of CzEng. Both treebanks follow the annotation style based on the Prague tectogrammatics theory (Sgall, 1967; Sgall et al., 1986). While PCEDT consists of 50,000 sentence pairs annotated mostly manually, the annotation of CzEng with 15 million parallel sentences is entirely automatic. Both treebanks have been provided with a fully automatic alignment of Czech and English nodes (Mareček et al., 2008), which is, however, prone to errors for *it* and its Czech counterparts. Since they are pronouns, they can replace a wide range of content words and their meaning is inferred mainly from the context. The situation is better for verbs as their usual parents in dependency trees: since they carry meaning in a greater extent, their automatic alignment is of a higher quality.

We took advantage of this property and the gold annotation of semantic roles in PCEDT, obtaining Czech translations as the argument of the Czech verb aligned with the English parent verb that fills the same semantic role as the given *it*. Using this approach, we succeeded in reaching the Czech counterpart in more than 60% of instances. The rest had to be done manually.

[3]`http://www.ldoceonline.com`

It	Train	Test	**Reflexives**	Train	Test
PCEDT sections	00–19	20–21	CzEng sections	00–09	98
PersPron	576	322	Se	6,305	652
To	231	138	Sam	2,271	205
Null	133	83	SamSe	1,361	129
			Samotny	804	89
Total	940	543	Total	10,741	1,075

Table 1: Distribution of classes in the data sets.

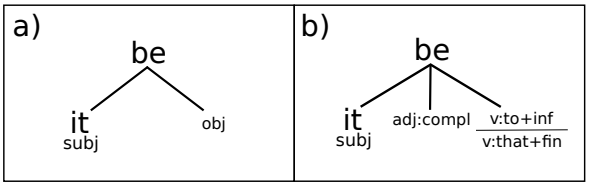

Figure 2: Examples of syntactic features capturing typical constructions with a verb *be*.

Czech counterparts of English reflexive pronouns have been collected directly from the alignment in CzEng, ignoring the cases where the aligned Czech word does not fall in one of the classes mentioned in Section 2.2.

The overall statistics of the train and the test set are shown in Table 1. The disproportion of training instances for *it* results from the manual annotation of classes, which could not be completely finished due to time reasons. In order to maintain the overall distribution, we also had to limit the number of automatically annotated classes.

Given the observation (see Section 2), we designed features to differentiate between the ways *it* and reflexives are translated.

3.1 Features for *it*

The translation mapping in Figure 1 suggests that identifying the English type of *it* might be informative. We thus constructed a binary coreference-related feature based on the output of the system NADA (Bergsma and Yarowsky, 2011) giving an estimate of whether an instance of *it* is coreferential.

Some verbs are more likely to bind with *it* that refers to a longer utterance. Such *it* is relatively consistently translated as a demonstrative *to*. However, PCEDT is too small to be a sufficient sample from a distribution over lexical properties. Hence, we took advantage of CzEng and collected co-occurrence counts between a semantic role that the given *it* fills concatenated with a lemma of its verbal parent and a Czech counterpart having the same semantic role (denoted as $csit$). We filtered out all occurrences where $csit$ was neither a personal pronoun nor *to*. For both possible values of $csit$ a feature is constructed by looking up frequencies for a concrete occurrence in the co-occurrence counts collected on CzEng and quantized into 4-5 bins following the formula:

$$bin(\log(\frac{count(semrole : parent \wedge csit)}{count(semrole : parent)count(csit)})).$$

Linguistic analysis suggested including syntax-oriented patterns related to the verb *to be* such as those shown in Figure 2. For instance, nominal predicates[4] tend to be translated as *to* even if *it* is coreferential. On the other hand, an adjectival predicate followed by a subordinating clause introduced by the English connectives *to* or *that* usually indicates a pleonastic usage of *it* translated as a null subject.

3.2 Features for reflexive pronouns

Here we focused on distinguishing between the two most frequent meanings (see Section 2.2). Ideally, the POS tag of the parent would be a sufficient feature because reflexives in the second meaning should depend on a noun. However, since we deal with automatically parsed trees we had to support the parent POS tag by the POS tag of the immediately preceding word. Moreover, another feature indicates if the preceding word is a noun and agrees with the pronoun in gender and number.

Furthermore, we observed that *sám* rarely appears in other case than nominative. Although this feature exploits the target side, we can use it since the case of the governing Czech noun is already known at the point when reflexives are translated.

Last but not least, the morpho-syntactic pattern (including a possible preposition) in which the reflexive pronoun appears is a valuable feature.

4 Experiments and Evaluation

To mitigate a possible error caused by a wrong classifier choice, we built several models based on various Machine Learning classification methods including Maximum Entropy inplemented in the AI::MaxEntropy Perl library,[5] logistic regression with one-against-all strategy from Vowpal Wabbit[6] as well as decision trees, k-NN and SVM from Scikit-learn library (Pedregosa et al., 2011).

[4] The verb *to be* has an object.

[5] http://search.cpan.org/~laye/AI-MaxEntropy-0.20

[6] http://hunch.net/~vw

	It		Reflexives	
	Train	Test	Train	Test
Baseline	60.70	59.30	58.70	60.65
AI::MaxEntropy	85.99	**76.61**	76.37	77.77
VW (passes=20, l2=10e-5)	89.99	76.43	76.98	77.77
sklearn:decision-trees	93.36	73.66	81.78	76.37
sklearn:k-NN (k=10)	82.51	73.30	77.64	76.74
sklearn:SVM (kernel=linear)	90.83	75.51	76.55	**78.14**

Table 2: Accuracy of both translation models on the training and test data.

	It	Reflexives
new better than old	24	12
old better than new	13	0
equal quality	13	5

Table 3: The results of manual evaluation on sentences translated by TectoMT in the original settings and using the new translation models

We compare our results with a majority class baseline (`PersPron` and `Se` classes) in Table 2. The results show a 17% gain when our approach is used.

The specialized models have been integrated in the TectoMT system and extrinsically evaluated on the English-Czech test set for the WMT 2011 Translation Task (Callison-Burch et al., 2011).[7] This data set contains 3,003 English sentences with one Czech reference translation, out of which 430 contain at least one occurrence of *it* and 52 contain a reflexive pronoun.

The new approach was compared to the original TectoMT rule-based pronoun handling heuristics (see Section 1). The shift from the original settings to the new translation models results in 166 changed sentences with *it* and 17 changed sentences with English reflexives. In terms of BLEU score, we observe a marginal drop from 0.1404 to 0.1403 using the new approach. However, BLEU may be too coarse for this kind of experiment.

In order to give a more realistic view, we carried out a manual evaluation. All 17 modified sentences for reflexives and 50 randomly sampled changed sentences containing *it* were presented to one annotator who assessed which of the two systems gave a better translation. Table 3 shows that improved sentences dominate in both cases. Overall, the improved sentences account for around 8.5% of all sentences with *it* and 23% sentences containing a reflexive pronoun.

5 Discussion

Looking into the types of improvements and errors in the manually evaluated sentences, we have found that the new model for *it* opted for a different translation only in cases where the original system decided to express *to* on the surface. In 13 out of 24 improvements, the new model for *it* succeeded in correctly resolving the `Null` class while in the remaining 11 cases, the corrected class was `PersPron`. It took advantage mostly of the syntax-based features in the former case and the coreference-related feature in the latter.

Regarding the reflexive pronouns the pronoun was used in its emphasizing meaning in all but two altered sentences. This accords with the design of features, which are mainly targeted at revealing this usage of reflexives. Moreover, the feature indicating if a Czech noun is in nominative case has proved to be particularly useful, correctly driving the lexical choice between *sám* and *samotný*. The majority of errors stem from incorrect activation of syntactic features due to parsing and POS tagging errors.

6 Conclusions

In this work, we presented specialized translation models for two types of English pronouns: *it* and reflexives. Integrating them into an English-Czech syntax-based MT system TectoMT we succeeded in improving the concerned sentences measured by human evaluation.

Generally, it is intractable to design a specific feature set for every word. However, this work shows on two examples that the correct translation of some words depends on many linguistic aspects, e.g. syntax and coreference and that is worth taking these aspects into account.

Acknowledgments

This work has been supported by the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875), the grant GAUK 4226/2011 and EU FP7 project Khresmoi (contract no. 257528). This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

[7] `http://www.statmt.org/wmt11/test.tgz`

References

Shane Bergsma and David Yarowsky. 2011. NADA: A Robust System for Non-Referential Pronoun Detection. In *DAARC*, pages 12–23, Faro, Portugal, October.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, May. ELRA, European Language Resources Association.

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, Edinburgh, Scotland, July. Association for Computational Linguistics.

Liane Guillou. 2012. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the EACL*, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the NAACL HLT – Volume 1*, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Uppsala, Sweden, July. Association for Computational Linguistics.

David Mareček, Zdeněk Žabokrtský, and Václav Novák. 2008. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In *Proceedings of the Twelfth EAMT Conference*, pages 102–111.

Izak Morin. 2009. Translating Pronouns, Proper Names and Kinship Terms from Indonesian into English and vice versa. *TEFLIN Journal: A publication on the teaching and learning of English*, 16(2).

Kristýna Onderková. 2010. Possessive Pronouns in English and Czech Works of Fiction, Their Use with Parts of Human Body and Translation. Master's thesis, Masaryk University, Faculty of Arts.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, November.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Stroudsburg, PA, USA. Association for Computational Linguistics.