Translation of "It" in a Deep Syntax Framework

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

Abstract

We present a novel approach to the translation of the English personal pronoun *it* to Czech. We conduct a linguistic analysis on how the distinct categories of *it* are usually mapped to their Czech counterparts. Armed with these observations, we design a discriminative translation model of *it*, which is then integrated into the TectoMT deep syntax MT framework. Features in the model take advantage of rich syntactic annotation TectoMT is based on, external tools for anaphoricity resolution, lexical co-occurrence frequencies measured on a large parallel corpus and gold coreference annotation. Even though the new model for *it* exhibits no improvement in terms of BLEU, manual evaluation shows that it outperforms the original solution in 8.5% sentences containing *it*.

1 Introduction

After it has long been neglected, retaining cohesion of a text larger than a single sentence in Machine Translation (MT) has recently become a discussed topic. Correct translation of referential expressions is in many cases essential for humans to grasp the meaning of a translated text.

Especially, the translation of pronouns attracts a higher rate of interest. In the previous works of Le Nagard and Koehn (2010), Hardmeier and Federico (2010) and Guillou (2012), it has been shown that current MT systems perform poorly in producing the correct forms of pronouns. As regards English, the personal pronoun *it* is the most complicated case. Not only can it corefer with almost any noun phrase (making it hard to pick the correct gender and number if the target language is morphologically rich), but it can also corefer with a larger discourse segment or play the role of a filler in certain grammatical constructions.

In this work, we turn our attention to the translation of the English personal pronoun *it* into Czech. Even if we ignore morphology and merge all related surface forms into one, we cannot find a single Czech expression that would comprise all functions of the English *it*. Moreover, there is no simple one-to-one mapping from categories of *it* to Czech expressions. For instance, one would expect that the translation of *it* which is coreferential with a noun phrase has to agree in number and gender with the translation of its antecedent. However, there are cases when it is more suitable to translate *it* as the demonstrative pronoun *to*, whose gender is always neuter.

The aim of this work is to build an English-to-Czech translation model for the personal pronoun *it* within the TectoMT framework (Žabokrtský et al., 2008). TectoMT is a tree-to-tree translation system with transfer via tectogrammatical layer, a deep syntactic layer which follows the Prague tectogrammatics theory (Sgall, 1967; Sgall et al., 1986) Therefore, its translation model outputs the deep syntactic representation of a Czech expression. Selecting the correct grammatical categories and thus producing a concrete surface form of a deep syntactic representation is provided by the translation synthesis stage, which we do not focus on in this work.

The mapping between *it* and corresponding Czech expressions depends on many aspects. We address them by introducing features based on syntactic annotation and anaphoricity resolver output. Furthermore, we make use of lexical co-occurrence counts aggregated on a large automatically annotated Czech-English parallel corpus CzEng 1.0 (Bojar et al., 2012). Coreference links also appear to be a source of valuable features.[1]

In contrast to the related work, we prefer a discriminative model to a commonly used generative

[1]However, we excluded them from the final model used in MT as they originate from gold standard annotation.

Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT), pages 51–59,
Sofia, Bulgaria, August 9, 2013.

model. The former allows us to feed it with many syntactic and lexical features that may affect the output, which would hardly be possible in the latter.

2 Related Work

Our work addresses a similar issue that has been explored by Le Nagard and Koehn (2010), Hardmeier and Federico (2010) and Guillou (2012). These works attempted to incorporate information on coreference relations into MT, aiming to improve the translation of English pronouns into morphologically richer languages. The poor results in the first two works were mainly due to imperfect automatic coreference annotation.

The work of Guillou (2012) is of special interest to this work because it is also focused on English to Czech translation and makes an extensive use of the Prague Czech-English Dependency Treebank 2.0 (PCEDT). Instead of automatic coreference links, they employed gold annotation, revealing further reasons of small improvements – the number of occurrences in the tranining data weakened by including grammatical number and gender in the annotation and availability of only a single reference translation.

The first issue is a consequence of the assumption that a Czech pronoun must agree in gender and number with its antecedent. There are cases, though, when demonstrative pronoun *to* fits better and grammatical categories are not propagated. Keeping grammatical information on its antecedent may in this case result in probably not harmful but still superfluous partitioning the training data.

Our work deals also with the second issue, however, at the cost of partial manual annotating.

The most significant difference of our work compared to the abovementioned ones lies in the MT systems used. Whereas they tackle the issue of pronoun translation within the Moses phrase-based system (Koehn et al., 2003), we rely on the translation via deep syntax with TectoMT system (Žabokrtský et al., 2008). Our approach is more linguistically oriented, working with deep syntactic representations and postponing the decisions about the concrete forms to the synthesis stage.

3 Linguistic Analysis

In English, three main coarse-grained types of *it* are traditionally distinguished. Referential *it* points to a noun phrase in the preceding or the following context:

(1) Peter has finished writing an article and showed *it* to his supervisor.

Anaphoric *it* refers to a verbal phrase or larger discourse segments (so-called discourse deixis).

(2) Peter has discussed the issue with his supervisor and *it* helped him to finish the article.

Pleonastic *it* has no antecedent in the preceding/following context and its presence is imposed only by the syntactic rules of English.

(3) *It* is difficult to give a good example.

From the perspective of Czech, there are also three prevailing types of how *it* can be translated. The most frequent are personal pronouns or zero forms.[2] In Prague tectogrammatics theory zero anaphors are reconstructed on the tectogrammatical layer. Same as expressed personal pronouns, they are represented by a node with the *#PersPron* symbol, e.g.

(4) Bushova vláda oznámila, že se svůj plán *#PersPron* pokusí vzkřísit.

The Bush administration has said *it* will try to resurrect its plan.

The second typical possibility is the Czech demonstrative pronoun *to* (= it, this), which is a form of a pronoun *ten* in its neuter singular form, e.g.

(5) Analytik řekl, že *to* byla tato možnost požadavku, která pevnějším cenám pomohla.

The analyst said that *it* was the possibility of this demand that helped firm prices.

In many cases, it has no lexical counterpart in the Czech translation, the English and Czech sentences thus having a different syntactic structure. These are cases like, for instance:

(6) Obchodníci uvedli, že *je obtížné* nové emise REMIC strukturovat, když se ceny tolik mění.

Dealers noted that *it's difficult* to structure new Remics when prices are moving widely.

[2] Czech is a pro-drop language.

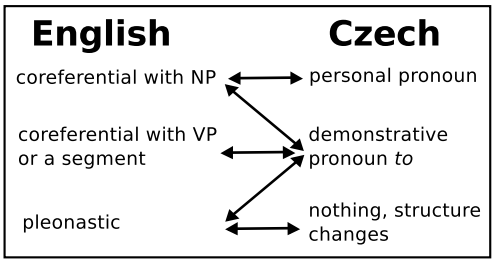

Figure 1: The mapping of the types of English *it* to Czech translations.

There are also some other possibilities of how *it* can be translated into Czech, such as the repetition of the antecedent noun, different genders of the demonstrative *ten* (=*it*, *this*) in the anaphoric position, using synonyms and hyperomyns. However, these cases are not so frequent and they rarely cannot be converted to one of the three broader categories.

The correspondence between the course-grained types of English *it* and its possible Czech translations is not one-to-one. As seen from Figure 1, a personal pronoun/zero anaphora translates to the referential *it* (see example 4) and no lexical counterpart is used when translating the pleonastic it (see example 6).

However, all types of *it* can be translated as a neuter demonstrative *to*. The typical case "*it* referring to VPs/larger discourse segments = *to*" was demonstrated in (5).

The mapping "referential *it* = *to*" is common for cases where the referent is attributed some further characteristics, mostly in constructions with a verb *to be* like "It is something.", such as (7).[3] This is an interesting case for Czech, because a gender and number agreement between the antecedent and the anaphoric *to* is generally absent.

(7) 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.

The "cleft sentences" (see example 8) and some other syntactic constructions are the case when pleonastic *it* is translated into Czech with the demonstrative *to*.

[3]We suspect that it holds also for *he*/*she*/*they* but such a claim is not yet empirically supported. For the sake of simplicity, we conduct our research only for *it*.

(8) But *it* is Mr. Lane, as movie director, who has been obsessed with refitting Chaplin's Little Tramp in a contemporary way.

Ale je *to* Lane jako filmový režisér, kdo je posedlý tím, že zmodernizuje Chaplinův film "Little Tramp (Malý tulák)".

In some cases, both translations of pleonastic *it* are possible: neuter demonstrative *to* or a different syntactic construction with no lexical counterpart of *it*. Compare the examples from PCEDT where *it* with similar syntactic function was translated by changing the syntactic structure in (9) and using a neuter *to* in (10):

(9) "*It was great* to have the luxury of time," Mr. Rawls said.

"*Bylo skvělé*, že jsme měli dostatek času," řekl Rawls.

(10) "On days that I'm really busy," says Ms. Foster, "*it seems decadent* to take time off for a massage."

"Ve dnech, kdy mám opravdu mnoho práce," říká paní Fosterová, "*to vypadá zvrhle*, když si vyhradím čas na masáž."

4 Translation via Deep Syntax

Following a phrase-based statistical MT approach, it may be demanding to tackle issues that arise when translating between typologically different languages. Translation from English to Czech is a typical example. One has to deal with a rich morphology, less constrained word order, changes in clauses bindings, pro-drops etc.

In this work, we make use of the English to Czech translation implemented within the TectoMT system, first introduced by Žabokrtský et al. (2008). In contrast to the phrase-based approach, TectoMT performs a tree-to-tree machine translation. Given an input English sentence, the translation process is divided into three stages: analysis, transfer and synthesis. TectoMT at first conducts an automatic analysis including POS tagging, named entity recognition, syntactic parsing, semantic role labeling, coreference resolution etc. This results in a deep syntactic representation of the English sentence, which is subsequently transferred into Czech, with the translation of lexical and grammatical information being provided via several factors. The process proceeds with a rule-

based synthesis stage, when a surface Czech sentence is generated from its deep syntactic structure.

Deep syntactic representation of a sentence follows the Prague tectogrammatics theory (Sgall, 1967; Sgall et al., 1986). It is a dependency tree whose nodes correspond to the content words in the sentence. Personal pronouns missing on the surface are reconstructed in special nodes. Nodes are assigned semantic roles (called functors) and grammatical information is comprised in so called grammatemes. Furthermore, tectogrammatical representation is a place where coreference relations are annotated.

4.1 Model of *it* within TectoMT

The transfer stage, which maps an English tectogrammatical tree to a Czech one, is a place where the translation model of *it* is applied. For every English node corresponding to *it*, a feature vector is extracted and fed into a discriminative resolver that assigns one of the three classes to it – `PersPron`, `To` and `Null`, corresponding to the main Czech types introduced in Section 3.

If labeled as `PersPron`, the English node is mapped to a Czech *#PersPron* node and the English coreference link is projected. During the synthesis, it is decided whether the pronoun should be expressed on a surface, its gender and number are copied from the antecedent's head and finally the correct form (if any) is generated.

Obtaining class `To` makes things easier. The English node is only mapped to a Czech node containing the pronoun *ten* with its gender and number set to neuter singular, so that later the correct form *to* will be generated.

Last, if *it* is assigned `Null`, no corresponding node on the Czech side is generated, but the Czech counterpart of the governing verb is forced to be in neuter singular.

5 Prague Czech-English Dependency Treebank as a source of data

The Prague Czech-English Dependency Treebank (Hajič et al., 2011, PCEDT) is a manually parsed Czech-English parallel corpus comprising over 1.2 million words for each language in almost 50,000 sentence pairs. The English part contains the entire Penn Treebank–Wall Street Journal Section (Linguistic Data Consortium, 1999). The Czech part consists of translations of all the texts from the English part. The data from both parts are annotated on three layers following the theory of Prague tectogrammatics – the morphological layer (where each token from the sentence gets a lemma and a POS tag), the analytical layer (surface syntax in the form of a dependency tree, where each node corresponds to a token in the sentence) and the tectogrammatical representation (see Section 4).

Sentences of PCEDT have been automatically morphologically annotated and parsed into analytical dependency trees.[4] The tectogrammatical trees in both language parts have been annotated manually (Hajič et al., 2012). The nodes of Czech and English trees have been automatically aligned on analytical as well as tectogrammatical layer (Mareček et al., 2008).

5.1 Extraction of Classes

The shortcomings of the automatic alignment is particularly harmful for pronouns and zero anaphors, which can replace a whole 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.

Thus, we did not search for a Czech counterpart of *it* by following the alignment of *it* itself. Using the fact that the verb alignment is more reliable and functors in tectogrammatical trees have been manually corrected, we followed the alignment of the parent of *it* (a verb) and selected the Czech subtree with the same tectogrammatical functor as *it* had on the English side. If the obtained subtree is a single node of type *#PersPron* or *ten*, we assigned class `PersPron` or `To`, respectively, to the corresponding *it*. This approach relies also on the assumption that semantic roles do not change in the translation.

The automatic acquisition of classes covered more than 60% of instances, the rest had to be labeled manually. During the annotation, we obeyed the following rules:

1. If a demonstrative pronoun *to* is present in the Czech sentence or if a personal pronoun is either present or unexpressed, assign the instance to the corresponding class.

[4]The English dependency trees were built by automatically transforming the original phrase-structure annotation of the Penn Treebank.

2. Otherwise, ignore the Czech translation provided in the corpus and follow the most simplistic possible translation which would still be correct. Assign the instance to the class which fits it the best.

Note that it may happen that none of the three options fits, because it is either an idiomatic expression or larger structural modifications are required. Such cases are very rare and we left them out of the data.

The manual annotation was a bottleneck. We managed to tag the complete testing data, but were only able to annotate more than just 1/6 of the training data due to time reasons. We only use a corresponding proportion of the automatically labeled training instances in order to respect the overall distribution.

5.2 Extraction of Features

Given the linguistically supported observation on both manually and automatically annotated treebanks, we designed features to differentiate between the ways *it* is translated.

Since this work focuses on MT with transfer via deep-syntactic layer, it is possible for the proposed features to exploit morphological, syntactic and a little of semantic information present on various annotation layers.

Unlike the target classes, which have to be assigned as accurately as possible, extracted features must follow the real-world scenario of MT – the only information that is given is the source sentence. Thus, whereas extracting classes may exploit the gold standard linguistic annotation, it cannot be employed in feature extraction. We extract them from text automatically annotated by the same pipeline that is used in the TectoMT analysis stage.

However, there is an exception where we violate this approach – coreference. Performance of state-of-the-art coreference resolvers is still far from the ideal, especially for distinguishing between pronouns referring to noun phrases and those referring to clauses or wider discourse segments. Similarly to the work of Guillou (2012) we wanted to isolate the problem of translating referential expressions from the task of resolving the entity they refer to. Therefore, we opted for extracting the coreferential features from the gold annotation projected onto automatically analyzed trees. Note that the results achieved using these features have to be considered an upper bound for a given setting.

Although the mapping between Czech translation of *it* and English categories of *it* does not allow to translate *it* directly, the category of *it* estimated by an anaphoricity resolver might be a promising feature. We therefore constructed a binary feature based on the output of a system identifying whether a pronoun *it* is coreferential or not. We employed the NADA resolver (Bergsma and Yarowsky, 2011)[5] exploiting the web-scale n-gram data and its tree-based extension presented in (Veselovská et al., 2012).

Some verbs are more likely to bind with *it* that refers to a longer utterance. Such *it* is quite consistently translated as a demonstrative *to*. This motivated incorporating a parent lemma of an occurrence of *it* into the feature set. However, the training data is too small to be a sufficient sample from a distribution over lexical properties. Hence, we took advantage of the automatically annotated[6] Czech-English corpus CzEng 1.0 (Bojar et al., 2012) that comprises more than 15 million sentence pairs. In the manner described in Section 5.1, we collected co-occurrence counts between a functor that the given *it* possesses concatenated with a lemma of its verbal parent and a Czech counterpart having the same functor (denoted as $csit$). We filtered out all occurrences where $csit$ was neither *#PersPron* nor *ten*. Then, for both values of $csit$ a feature is constructed by looking up counts for a concrete occurrence in the collected counts and quantized into 4-5 bins (Bansal and Klein, 2012) following the formula:

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

Linguistic analysis carried out in Section 3 suggests the following syntax-oriented features related to the verb *to be*. Some nominal predicates tend to be translated as *to*, even though *it* is usually coreferential in such expressions (see example 7). So the corresponding binary feature fires if *it* is a subject and its parent is the verb *to be* having an object (Figure 2a).

Similarly, adjectival predicates that are not followed by a subordinating clause connected with

[5] A probability value returned by this tool was binarized at a threshold 0.5

[6] Using the same annotation layers as in PCEDT and TectoMT, i.e. in accordance with the Prague tectogrammatics theory.

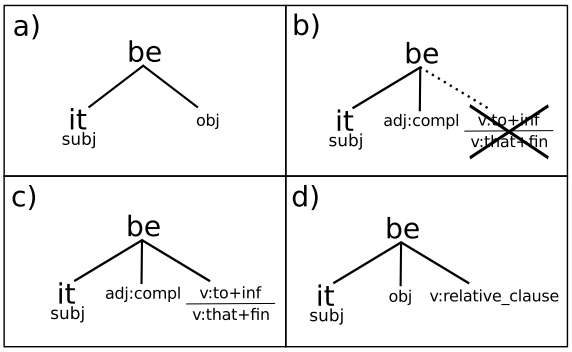

Figure 2: Syntactic features capturing typical constructions with a verb *be*.

the main clause by the English connectives *to* or *that* are usually referential and translated as *to*, too. We proposed a feature describing these cases, illustrated in Figure 2b.

In contrast, if an adjectival predicate is followed by a subordinating clause with the verb being finite and connected to the main clause by a conjunction *that*, in majority of cases it is a pleonastic usage of *it* translated as a null subject (see example 6). A schema of the feature is depicted in Figure 2c.

Being definitely pleonastic, *it* in cleft sentences is expressed in Czech either by *to* or by sentence rearranging (see example 8). We target this phenomenon by another feature being fired if *it* is a subject of the verb *to be* and if this verb has an object and is followed by a relative clause (see Figure 2d).

Finally, we designed two features exploiting coreference relations. The first one simply indicates if *it* has an antecedent, while the second fires if any of the antecedents in the coreferential chain is a verb phrase. As we noted above, these features are based on the gold standard annotation of coreference.

5.3 Data Description

The data for training and testing a discriminative translation model of the personal pronoun *it* were extracted from PCEDT with classes and features obtained as described in Section 5.1 and 5.2, respectively. Due to the limited amount of manually annotated training data, the training set extracted from sections 00 – 19 was reduced from 5841 to 940 instances, though. The testing set was annotated thoroughly, thus containing 543 instances extracted from sections 20 – 21. Every instance represents an occurrence of *it* in PCEDT. The distribution of target classes in the data is shown in Table 1.

Class	Train	Test
`PersPron`	576	322
`To`	231	138
`Null`	133	83

Table 1: Distribution of classes in the data sets.

6 Experiments

Experiments were conducted in two settings that differ in the usage of features extracted from gold coreferential relations.

To mitigate a possible error caused by a wrong classifier choice, we built several models based on various Machine Learning classification methods. If not explicitly mentioned, the methods below are applied with default parameters:

- **Vowpal Wabbit** (Langford, 2012). Binary logistic regression with one-against-all strategy for handling multiple classes. The optimum has been found using the online method (Stochastic Gradient Descent). We varied the parameters of the number of passes over the data and the L2 regularization weight.
- **AI::MaxEntropy**.[7] Multiclass logistic regression.[8] The optimum has been found using the batch method (L-BFGS).
- **sklearn.neighbors**.[9] `k`-nearest neighbors classifier with the parameter `k` being varied.
- **sklearn.tree**. Decision tree classifier.
- **sklearn.SVC**. Support Vector Machines with one-against-one strategy to handle multiple classes. We varied the choice of a kernel.

The accuracy evaluated on both training and test sets is shown in Table 2 (columns Acc:Train and Acc:Test). The baseline resolver simply picks the most frequent class in the training set, which is `PersPron`. For both experimental settings, the standard deviation measured on the test set is less than 1% in total, if the method's best configuration of parameters is taken and the result on decision trees, which we did not tune, is excluded. This shows that all classifiers are consistent in their decisions.

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

[8] In the field of NLP also called Maximum Entropy.

[9] All classifiers labeled as sklearn.* are implemented in the Scikit-learn Python library (Pedregosa et al., 2011).

ML Method	all feats Acc:Train	all feats Acc:Test	all feats BLEU	all feats + coref Acc:Train	all feats + coref Acc:Test
Baseline	60.70	59.30	0.1401	60.70	59.30
Original TectoMT	–	–	**0.1404**	–	–
Vowpal Wabbit (passes=30)	90.62	75.69	–	90.83	75.87
Vowpal Wabbit (passes=20)	89.99	76.43	0.1403	90.20	76.98
Vowpal Wabbit (passes=10)	87.78	76.24	–	87.78	76.61
Vowpal Wabbit (passes=30, l2=0.001)	71.23	66.11	–	83.03	**77.16**
Vowpal Wabbit (passes=20, l2=0.001)	82.19	74.95	–	78.19	74.40
Vowpal Wabbit (passes=10, l2=0.001)	75.03	70.17	–	72.81	70.17
Vowpal Wabbit (passes=30, l2=0.00001)	90.52	75.69	–	90.94	76.06
Vowpal Wabbit (passes=20, l2=0.00001)	89.99	76.43	–	90.09	76.98
Vowpal Wabbit (passes=10, l2=0.00001)	87.67	76.24	–	87.67	76.61
AI::MaxEntropy	85.99	**76.61**	0.1403	86.09	76.98
sklearn.neighbors (k=1)	91.57	71.64	–	93.36	72.19
sklearn.neighbors (k=3)	84.62	72.01	–	84.93	71.82
sklearn.neighbors (k=5)	84.93	74.77	0.1403	84.72	75.87
sklearn.neighbors (k=10)	82.51	73.30	–	83.14	75.87
sklearn.tree	93.36	73.66	0.1403	94.10	71.82
sklearn.SVC (kernel=linear)	90.83	75.51	0.1402	91.15	76.80
sklearn.SVC (kernel=poly)	60.70	59.30	–	60.70	59.30
sklearn.SVC (kernel=rbf)	71.23	68.69	–	73.76	71.27

Table 2: Intrinsic (accuracy on the training and test data) and extrinsic (BLEU score) evaluation of translation model of *it* in configuration with (all feats) and without gold coreferential features (all feats + coref).

By introducing linguistically motivated features exploiting the deep-syntactic description of the sentence, we gained 17% in total over the baseline. Moreover, adding features based on the gold coreference annotation results in a further 0.5% improvement.

7 Evaluation on MT

Although intrinsic evaluation as performed in Section 6 can give us a picture of how accurate the translation model might be, the main purpose of this work is to integrate it in a full-fledged MT system. As explained in Section 4, this component is tailored for TectoMT – an MT system where the transfer is provided through a deep-syntactic layer.

The extrinsic evaluation of the proposed method was carried out on the English-Czech test set for WMT 2011 Shared Translation Task (Callison-Burch et al., 2011).[10] This data set contains 3,003 English sentences with one Czech reference translation, out of which 430 contain at least one occurrence of *it*.

Since this test set is provided with no annotation of coreferential links, the model of *it* that is involved in experiments on the end-to-end translation was trained on a complete feature set excluding the coreferential features using the Machine Learning method that performed best in the intrinsic test, i.e. AI::MaxEntropy (see Section 6).

The new method was compared to the rule-based approach originally used in TectoMT, which works as follows. In the transfer stage, all occurrences of *it* are translated to a demonstrative *ten*. In the synthesis stage, another rule is fired, which determines whether *ten* is omitted on the surface. Then, omitting it corresponds either to a structural change (`Null` class) or an unexpressed personal pronoun (a subset of `PersPron` class). It makes this original approach difficult to compare with the scores in Table 2, as the translation model of *it* is applied in the transfer stage, where we do not know yet if a personal pronoun is to be expressed or not. Thus, we consider it the most appropriate to use final translated sentences produced by two versions of TectoMT in order to compare the different way they handle *it*.

The shift from the original settings to a new model for *it* results in 166 changed sentences. In terms of BLEU score, we observe a marginal drop from 0.1404 to 0.1403 when using the new approach.[11] Other classifiers achieved the same or

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

[11] For comparison, the best system so far – Chimera (Bojar et al., 2013) achieves 0.1994 on the same test set. Chimera combines Moses, TectoMT and rule-based corrections.

new better than old	24
old better than new	13
both equally wrong	9
both equally correct	4

Table 3: The results of manual evaluation conducted on 50 sentences translated by TectoMT in the original settings (old) and with the new translation model for *it* (new)

similar score which correlates with the findings from intrinsic evaluation (see Table 2). It accords with a similar experience of Le Nagard and Koehn (2010) and Guillou (2012) and gives another evidence that the BLEU metric is inaccurate for measuring pronoun translation.

Manual evaluation gives a more realistic view. We randomly sampled 50 out of the 166 sentences that differ and one annotator assessed which of the two systems gave a better translation. Table 3 shows that in almost half of the cases the change was an improvement. Including the sentences that are acceptable for both settings, the new approach picked the correct Czech counterpart of *it* in 22% more sentences than the original approach. Since the proportion of the changed sentences accounts for almost 39% of all sentences containing *it*, the overall proportion of improved sentences with *it* is around 8.5% in total.

8 Discussion

Inspecting the manually evaluated translation for types of improvements and losses, we have found that in none of the changed sentences the original system decided to omit *ten* (obtained by the rule) on the surface. It shows that the new approach agrees with the original one on the way of omitting personal pronouns and mainly addresses the overly simplistic assignment of the demonstrative *ten*.

The distribution of target classes over corrected sentences is almost uniform. In 13 out of 24 improvements, the new system 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 and suggestions given by the NADA anaphoricity resolver in the latter.

Examining the errors, we observed that the majority of them are incurred in the structures with "it is". These errors stem mostly from incorrect activation of syntactic features due to parsing and POS tagging errors. Example 11 (the Czech sentence is an MT output) shows the latter, when the POS tagger erroneously labeled the word *soy* as an adjective. That resulted in activating the feature for adjectival predicates followed by *that* (Figure 2c) instead of a feature indicating cleft structures (Figure 2d), thus preferring the label `Null` to the correct `To`.

(11) SOURCE: *It is* just soy that all well-known manufacturers use now.

TECTOMT: *Je* ~~*to*~~ jen sójové, že známí výrobci všech používají teď.

9 Conclusion

In this work we presented a novel approach to dealing with the translation of the English personal pronoun *it*. We have shown that the mapping between the categories of *it* and the ways of translating it to Czech is not one-to-one. In order to deal with this, we designed a discriminative translation model of *it* for the TectoMT deep syntax MT framework.

We have built a system that outperforms its predecessor in 8.5% sentences containing *it*, taking advantage of the features based on rich syntactic annotation the MT system provides, external tools for anaphoricity resolution and features capturing lexical co-occurrence in a massive parallel corpus,

The main bottleneck that hampered bigger improvements is the manual annotation of the training data. We managed to accomplish it just on 1/6 of the data, which did not provide sufficient evidence for some specific features.

Our main objective of the future work is thus to reduce a need for manual annotation by discovering ways of automatic extraction of reliable classes from a semi-manually annotated corpus such as PCEDT.

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).

References

Mohit Bansal and Dan Klein. 2012. Coreference Semantics from Web Features. In *Proceedings of the 50th Annual Meeting of the ACL: Long Papers – Volume 1*, pages 389–398, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. 2013. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*. Under review.

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.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160. ELRA.

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.

John Langford. 2012. Vowpal Wabbit.

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.

Linguistic Data Consortium. 1999. Penn Treebank 3. LDC99T42.

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.

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.

Kateřina Veselovská, Giang Linh Nguy, and Michal Novák. 2012. Using Czech-English Parallel Corpora in Automatic Identification of It. In *The Fifth Workshop on Building and Using Comparable Corpora*, pages 112–120.

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.

Feature Weight Optimization for Discourse-Level SMT

Sara Stymne, Christian Hardmeier, Jörg Tiedemann and Joakim Nivre
Uppsala University
Department of Linguistics and Philology
Box 635, 751 26 Uppsala, Sweden
firstname.lastname@lingfil.uu.se

Abstract

We present an approach to feature weight optimization for document-level decoding. This is an essential task for enabling future development of discourse-level statistical machine translation, as it allows easy integration of discourse features in the decoding process. We extend the framework of sentence-level feature weight optimization to the document-level. We show experimentally that we can get competitive and relatively stable results when using a standard set of features, and that this framework also allows us to optimize document-level features, which can be used to model discourse phenomena.

1 Introduction

Discourse has largely been ignored in traditional machine translation (MT). Typically each sentence has been translated in isolation, essentially yielding translations that are bags of sentences. It is well known from translation studies, however, that discourse is important in order to achieve good translations of documents (Hatim and Mason, 1990). Most attempts to address discourse-level issues for statistical machine translation (SMT) have had to resort to solutions such as post-processing to address lexical cohesion (Carpuat, 2009) or two-step translation to address pronoun anaphora (Le Nagard and Koehn, 2010). Recently, however, we presented Docent (Hardmeier et al., 2012; Hardmeier et al., 2013), a decoder based on local search that translates full documents. So far this decoder has not included a feature weight optimization framework. However, feature weight optimization, or tuning, is important for any modern SMT decoder to achieve a good translation performance.

In previous research with Docent, we used grid search to find weights for document-level features while base features were optimized using standard sentence-level techniques. This approach is impractical since many values for the extra features have to be tried, and, more importantly, it might not give the same level of performance as jointly optimizing all parameters. Principled feature weight optimization is thus essential for researchers that want to use document-level features to model discourse phenomena such as anaphora, discourse connectives, and lexical consistency. In this paper, we therefore propose an approach that supports discourse-wide features in document-level decoding by adapting existing frameworks for sentence-level optimization. Furthermore, we include a thorough empirical investigation of this approach.

2 Discourse-Level SMT

Traditional SMT systems translate texts sentence by sentence, assuming independence between sentences. This assumption allows efficient algorithms based on dynamic programming for exploring a large search space (Och et al., 2001). Because of the dynamic programming assumptions it is hard to directly include discourse-level features into a traditional SMT decoder. Nevertheless, there have been several attempts to integrate intersentential and long distance models for discourse-level phenomena into standard decoders, usually as ad-hoc additions to standard models, addressing a single phenomenon.

Several studies have tried to improve pronoun anaphora by adding information about the antecedent, either by using two-step decoding (Le Nagard and Koehn, 2010; Guillou, 2012) or by extracting information from previously translated sentences (Hardmeier and Federico, 2010), unfortunately without any convincing results. To address the translation of discourse connectives, source-side pre-processing has been used to annotate surface forms either in the corpus or in the

Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT), pages 60–69,
Sofia, Bulgaria, August 9, 2013.