Extracting Translation Verb Frames*

Ondřej Bojar and Jan Hajič
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Praha 1, CZ-118 00, Czech Republic
{bojar,hajic}@ufal.mff.cuni.cz

Abstract

We describe a method for extracting translation verb frames (parallel subcategorization frames) from a parallel dependency treebank. The extracted frames constitute an important part of machine translation dictionary for a structural machine translation system. We evaluate our method independently, using a manually annotated test dataset, and conclude that the bottleneck of the method lies in quality of automatic word alignment of the training data.

1 Introduction

Structural machine translation (as opposed to statistical machine translation) is very sensitive to quality of translation dictionaries with respect to both annotation detail and domain specificity. Unfortunately, manual development or update of translation dictionaries is a very laborious task.

For our language pair, Czech and English, there are no machine translation dictionaries available. There are some machine readable dictionaries, such as (Svoboda 01) or the WinGED dictionary[1], but these were intended for human users and do not contain the required information either at all or in a very non-formalized fashion. Other researchers experimenting with structural Czech-to-English MT, such as (Čmejrek *et al.* 03), had to use very limited dictionaries containing single word translations only, too.

We aim at building a Czech-to-English machine translation dictionary with more detailed syntactic information. In particular, we need to support an English adaptation of the Ruslan MT system (Hajič 87). To begin with, the system requires a detailed knowledge of verb frames and their translations.

We use Prague Czech-English Dependency Treebank (PCEDT, (Čmejrek *et al.* 04)) to automatically extract translation verb frames to populate the dictionary with (domain specific) syntactic constructions.

The task of subcategorization frames extraction based on corpus or treebank data has been extensively studied in the past. Please see (Korhonen 02) or (Zeman & Sarkar 00) for more details and also detailed comparison of various methods by several authors. However, our goal is different from all cited authors. Instead of learning whether a modification of a verb is a valid member of the verb's subcategorization frame (i.e. a COMPLEMENT) or whether it can accompany any verb (i.e. an ADJUNCT), we need to answer a different question: "What type of modification, i.e. what surface form, preposition etc., should a translation system use in English given the Czech verb and the Czech modification type?"

2 Method

We use the term TRANSLATION VERB FRAME to denote a pair of Czech and English verbs accompanied by a list of pairs Czech modification-English modification (called MODIFICATION PAIRS). This definition implies that the number of "slots" in both English and Czech frames must be equal. The modification pairs encode core morphological and syntactic information, such as the preposition and the case. This is an example of a translation verb frame with one modification pair:

- *dělit=divide na+accusative=into*

For automatic extraction of translation verb frames, we need parallel dependency corpus aligned on the word level. In our case, PCEDT suits well for the task but word-to-word alignments had to be added. We employed the GIZA++ toolkit ((Och & Ney 03)), although it has never been adapted for our specific language pair.

*This work has been supported by the Ministry of Education of the Czech Republic Project No. LC536, the Grant Agency of the Czech Republic grants Nos. 405/03/0914 nad 201/05/H014, and the Grant Agency of the Charles University grant No. 351/2005.

[1] http://www.rewin.cz/

2.1 Observing Translation Frames

In the first step, OBSERVED TRANSLATION FRAMES are obtained in the following manner: all occurrences of Czech verbs are scanned and if the automatic GIZA++ alignment linked the Czech verb to an English one, the modifications of the verbs are matched to one another. To each Czech modification, an English modification is assigned such that there is a maximum of alignment edges linking the subtrees of the modifications. Obviously, this rather simplistic approach has some limitations; see Section 4 for more details.

2.2 Clean-up of Observed Frames

The second step deals with cleaning up and boosting the statistics of observed frames. We experimented with several different techniques, including combinations of them:

- No cleanup (marked as `raw`): Observed frames are used directly.
- Removal of low-frequent modification types (`freq`): All observed frames are simplified (reduced in size) by removing modification pair types that were not observed frequently enough in the whole set of observations (regardless the verb).
- Removal of badly aligned sentences (`giza`): GIZA++ provides each sentence with a measure of alignment confidence. We employ this measure to collect observed frames only from sentences with less problematic alignments.
- Only "very simple" Czech sentences (`vss`): We employ a rule based system described in (Bojar 03) to remove all sentence pairs where the Czech sentence has too complex structure or high risk of wrong automatic syntactic analysis with respect to verb modifications. As demonstrated in (Bojar 03), this procedure significantly improves parsing accuracy (at the cost of reducing available data size). Because the Czech sentences of PCEDT are parsed automatically, this data selection should improve the quality of observed translation frames, too.

2.3 Optional Statistical Filtering

As a third step, we optionally employ statistical filtering in order to further simplify the set of cleaned-up observed frames.

So far, we have experimented with one possible method of this filtering. The algorithm Apriori ((Agrawal *et al.* 93)) was designed to support sales: given a list of transactions (sets of items purchased), Apriori identifies typical relations such as: "Someone who buys bread buys often butter, too." Alternatively, the output of Apriori can be interpreted as a list of most common subsets of transactions. The application of Apriori in our situation is straightforward: every (cleaned-up) observed frame corresponds to a transaction and every modification pair corresponds to an item. The sub-transactions (subsets of translation frames) suggested by Apriori are collected to the dictionary.

As another option of this filtering, we could use one of the methods described in (Zeman & Sarkar 00) to automatically identify modifications that are typical for the verb (i.e. complements). The typical modifications and their translations should be listed in the dictionary (typical modifications have typical translations), while the translations of the adjuncts can be stored for all verbs together.

3 Evaluation

In order to evaluate the performance of the described collection and filtering methods, we prepared a small corpus of 140 sentences containing 400 occurrences of 200 different verbs. In total, the corpus contains 1005 verb modifications that were manually aligned to their English counterparts.

3.1 Evaluation Algorithms

The complete MT system is still under development, so we implemented three simple algorithms that assign an English translation to every modification of the Czech verb, given the Czech verb and all its modifications.[2] All the algorithms make use of the same version of the extracted dictionary (i.e. the same set of cleaned-up and filtered translation frames), the difference lies in the method of constructing the English frame:

- Algorithm A – Translation slot by slot regardless of the verb. This version of the algorithm first collects all the Czech slots and

[2] Obviously, there might be more correct translations of the Czech sentences, so different English verbs with different modification forms can be used. For the sake of simplicity, we do not take this into account and use only one reference translation.

Clean-up	Apriori	Algorithm	F-score	Precision	Recall	
giza	Apriori	A	68.4	52.9	96.7	← the best F-score
giza	Apriori	CBA	66.4	50.5	96.7	
giza	Apriori	BA	66.1	50.2	96.7	
giza	No Apriori	BA	66.1	49.8	98.0	
raw	No Apriori	A	58.2	41.3	98.9	← baseline
freq	Apriori	BA	56.5	52.9	60.6	← the best precision
vss	Apriori	BA	55.9	41.4	86.3	← the best of vss
vss	Apriori	C	30.0	33.5	27.2	← the worst result

Table 1: Precision and recall of various algorithms for translating verb modification types from Czech to English.

their English counterparts from the whole dictionary. When translating, it processes Czech observed frame and slot by slot chooses the most probable English translation.

- Algorithm B – Translation slot by slot taking the verb into account. Similarly to A, B operates on every observed slot individually. During the preprocessing phase, B collects the most probable translations of Czech slot for each Czech verb independently. If a an unseen combination of a Czech verb and a Czech slot comes up in the test data, we optionally employ the algorithm A as a back-off. We use the label "BA" to indicate that B was used with the back-off.

- Algorithm C – Translation according to best matching frame. Given a Czech observed frame, C searches among all known frames for the given verb. A winner frame (or occasionally more than one frame if the results are equal) is selected based on the size of intersection of the known and observed Czech slots. The translations are then chosen slot by slot scanning the winner frames only. If some of the observed slots are not listed in any of the winner frames, B or optionally A can be used to find a more general solution. We label these mixed algorithms CB and CBA.

3.2 Results

Table 1 summarizes the indicators of prediction quality of selected combinations of filtering and algorithms. PRECISION is the percentage of correctly chosen translations from the total number of translations produced. (The algorithms may offer several answers at once, if there is no clear winner.) RECALL is the percentage of slots where the method supplied at least one answer from the total of 1005 tested slots. The results are sorted descending by F-SCORE, the harmonic mean of precision and recall.

The best result (68.4 F-score) was achieved using the most simple method A trained on data from well-paired sentences only (giza) filtered by Apriori. The algorithms CBA and BA with the same training data are slightly lower in precision. As the fourth best result indicates, training on well-paired sentences and applying no further filtering is sufficient to achieve nearly the same performance. All these results are 18% better over the baseline which employed algorithm A trained directly on observed frames (no clean-up and no filtering).

The results document that the quality of word alignment has the greatest impact on the performance of verb frame translation. The filtering using Apriori contributes slightly but on the other hand, the best performing algorithm is the most simple one (A), probably due to limited training data size.

On the other hand, training only on sentences where the Czech sentence is "very simple" (vss) significantly decreases the performance. The reason probably lies in the fact that in simple sentences, the modifications of verbs are much shorter and we find less evidence for alignment of the Czech and English slots.

Our results on automatic extraction of translation verb frames are not directly comparable to any of the various methods cited by (Korhonen 02) mainly due to the different goal and also due to a different evaluation metric. (Korhonen 02) cites results by various authors with F-score ranging from 47 to 85. However, it is impor-

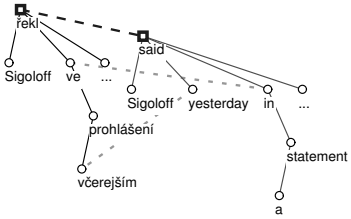

Figure 1: Sigoloff řekl ve včerejším prohlášení ... Sigoloff said yesterday in a statement ...

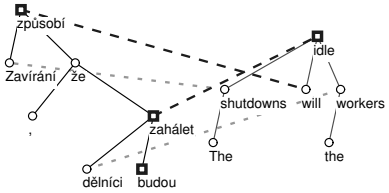

Figure 2: Zavírání způsobí , že dělníci budou zahálet ... The shutdowns will idle the workers ...

tant to note that her F-score is measured on the whole (monolingual) frames and not slotwise as in our evaluation. Only (Zeman & Sarkar 00) report slotwise F-score and reach 86 to 88. Most of the reported methods achieve more precise results than our approach, there is however a good reason for it: all the cited methods aim at answering a yes-no question only: "Is this modification a complement or and adjunct?" We aim at finding the correct English surface form given the Czech surface form of a modification. Naturally, the range of possible answers is much higher in our case.

4 Neglected Syntactic Divergences

For the sake of simplicity of our first experiments we neglect some important problems caused by syntactic differences between Czech and English as observed in our corpus of economical texts (PCEDT). In future research, we plan to investigate the divergences is more detail.

Modification shift. As illustrated in Figure 1, in some cases it is not possible to align verb modifications one-to-one due to a modification shift. Usually, part of speech is affected by the shift, too. An appropriate solution is to automatically identify these cases and ignore the shifted slots during training phase.

Head Switching. The preferred dependency analysis of some constructions such as modal verbs is different in Czech and in English, see Figure 3 for an instance. Most of these divergences are caused purely by different dependency annotation guidelines and can be handled accordingly by a set of rules.

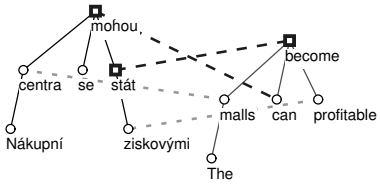

Figure 3: Nákupní centra se mohou stát ziskovými ... The malls can become profitable ...

However, there are situations where the divergence cannot be attributed to annotation guidelines only. (Dependency analysis by reduction, (Lopatková *et al.* 05), offers testable criteria and formal background for the distinction, however this type of analysis cannot be performed automatically, yet.) See Figure 2 for an example where English head verb was transformed to Czech sub-clause.

5 Conclusion and Further Research

We described an automatic procedure for extracting translation dictionary of verb frames from parallel word-aligned treebank. We evaluated various methods of data filtering and concluded that the quality of word-alignment is the bottleneck of the procedure. We also illustrated some syntactic divergences between Czech and English that should be handled with a special care.

In future research we plan to include proper handling of observed syntactic divergences either by adapting the non-parallel observations or at least by automatically identifying them to remove them from training data. With respect to the discovered bottleneck, we also plan to improve word-alignment performance for our language pair by taking more linguistic considerations into account.

References

(Agrawal *et al.* 93) Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, New York, NY, USA, 1993. ACM Press.

(Bojar 03) Ondřej Bojar. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120, 2003.

(Čmejrek *et al.* 03) Martin Čmejrek, Jan Cuřín, and Jiří Havelka. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90. Association for Computational Linguistics, April 2003.

(Čmejrek *et al.* 04) Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of LREC 2004*, Lisbon, May 26–28 2004.

(Hajič 87) Jan Hajič. RUSLAN: an MT system between closely related languages. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pages 113–117. Association for Computational Linguistics, 1987.

(Korhonen 02) Anna Korhonen. Subcategorization Acquisition. Technical Report UCAM-CL-TR-530, University of Cambridge, Computer Laboratory, Cambridge, UK, February 2002.

(Lopatková *et al.* 05) Markéta Lopatková, Martin Plátek, and Vladislav Kuboň. Modeling syntax of Free Word-Order Languages: Dependency Analysis By Reduction. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings*, volume LNAI 3658, pages 140–147. Springer Verlag, September 2005.

(Och & Ney 03) Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, 2003.

(Svoboda 01) Milan Svoboda. GNU/FDL English-Czech Dictionary, 2001. `http://slovnik.zcu.cz/`.

(Zeman & Sarkar 00) Daniel Zeman and Anoop Sarkar. Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000. ELRA.