Czech-English Word Alignment

Ondřej Bojar and Magdalena Prokopová

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague
bojar@ufal.mff.cuni.cz, magda.prokopova@gmail.com

Abstract

We describe an experiment with Czech-English word alignment. Half a thousand sentences were manually annotated by two annotators in parallel and the most frequent reasons for disagreement are described. We evaluate the accuracy of GIZA++ alignment toolkit on the data and identify that lemmatization of the Czech part can reduce alignment error to a half. Furthermore we document that about 38% of tokens difficult for GIZA++ were difficult for humans already.

1. Introduction

This article describes an experiment with Czech-English word alignment of the Prague Czech-English Dependency Treebank (PCEDT, (Čmejrek et al., 2004)).

Word alignment is usually used as a first step in the development of MT (machine translation) systems. A detailed understanding of the potential and the limits of word alignment between Czech and English is important for improving both SMT (statistical MT) as well as RBMT (rule-based MT) systems.

2. Corpus Data and Human Alignments

PCEDT 1.0 contains half of the Wall Street Journal section of Penn Treebank (21,000 sentences) translated sentence by sentence to Czech. The sentences are provided with surface and deep syntactic analyses, but no closer cross-language correspondence between the words or nodes is given.

In order to explore the problems of Czech-English word alignment, we conducted the following experiment: First, we observed a few sentences from PCEDT to agree upon simple annotation guidelines, similar to those proposed by (Melamed, 1998). We then prepared two independent manual annotations of 515 sentences.

Unlike previous approaches ((Martin et al., 2005), (Och and Ney, 2003) and others), who prepare golden standard alignments by discussing the problems and choosing a common solution, we pay more attention to the inter-annotator disagreement. We believe that the task of word alignment can be pursued only to a certain extent and that forcing full word alignment is not appropriate for all expressions. We use a simple rule of thumb to identify problematic cases: if our two annotators do not agree on the alignment of some words then the words deserve a more linguistically motivated analysis and possibly a slight redefinition of the word alignment task.

2.1. Sure, Possible and Phrasal Alignments

We asked the annotators to distinguish among cases where individual words match (SURE alignment), whole phrases correspond – but not words by themselves – (PHRASAL alignment) and cases when the connection is possible though doubtful (POSSIBLE alignment). The last case (possible connection) is used especially to connect words that do not have a real equivalent in the other language but syntactically clearly belong to a word nearby, such as English articles.

For phrasal alignments, annotators were encouraged to align also individual words in the phrases using sure or possible alignments, if reasonable.

The introduction of several types of connections allowed us to learn more about problematic cases (see below), but unfortunately phrasal alignments were used too rarely to draw any conclusions from them.

3. Inter-Annotator Agreement

As summarised in Table 1, data from our two annotators contain 32,000 connections all together (16,000 connections from each annotator on average). This includes two different alignments for each sentence. When we tried to evaluate inter-annotator mismatches we focused on connections that were made by one of the annotators and missing in the other annotator's data. This gave us 5,800 instances of possible problems, and inter-annotator mismatch reached 18%.

A closer look at situations where our annotators didn't agree revealed that half of these issues were caused by different selections of connection type. When the type of connection was omitted, the inter-annotator mismatch dropped to 9%.

Types of connections used		3	1
Annotator	A1	15,476	15,399
	A2	16,631	16,246
Mismatch	A1 but not A2	2,343	1,146
	A2 but not A1	3,498	1,714
Relative mismatch		18.2 %	9.0 %

Table 1: Inter-annotator mismatch.

In Table 2 we report for each word or part of speech the number of alignments where our annotators did not agree in connection type or the target word. A word aligned to one target word in one annotator's data and a different target word in the other annotator's data is counted twice.

Particular words that were most difficult for our annotators were articles, punctuation and prepositions even though when we compare part of speech categories and their misalignment we can see that nouns and verbs placed on the top of our scale in general. Single word misalignment demonstrates how complicated it is to align auxiliary words, words that don't carry meaning alone. These tokens can be

Problematic Words				Problematic Parts of Speech			
English		Czech		English		Czech	
361	to	319	,	679	IN	1348	N
259	the	271	se	519	DT	1283	V
159	of	146	v	510	NN	661	R
143	a	112	na	386	PRP	505	P
124	,	74	o	361	TO	448	Z
107	be	61	že	327	VB	398	A
99	it	55	.	310	JJ	280	D
95	that	47	a	245	RB	192	J
84	in	41	bude	216	NNP	59	C
80	by	37	k	199	VBN	22	T
...		...		...		...	

Table 2: Most frequent disagreeing alignments for words and parts of speech. **English Penn Treebank Tag-Set**: IN - Preposition or subordinating conjunction, DT - Determiner, NN - Noun, common, singular or mass, PRP - Pronoun, personal, TO - to, VB - Verb, base form, JJ - Adjective, NNP - Noun, proper, singular, VBN - Verb, past participle. **Czech Tag-Set**: N - Noun, V - Verb, R - Preposition, P - Pronoun, Z - Punctuation, sentence border, A - Adjective, D - Adverb, J - Conjunction, C - Number, T - Particle

frequently associated with several words when guidelines don't cover every single detail.

In the following, we describe the most common problems in a closer detail. For mentioned words we also report PERCENTAGE OF DISAGREEMENT, i.e. the number of misalignments divided by the number of all alignments of the token.

3.1. Articles

When dealing with languages that don't use articles in the same way or even worse, when one of these languages uses articles and the other one doesn't (like Czech and English) the word alignment task will generally generate a considerable amount of situations where it is difficult to connect these articles to a corresponding word. Unfortunately leaving articles out of alignment isn't the best solution since when the data is used for machine translation there must be way how to add them into the targeted text.

Evaluation of our data revealed that about one fifth of article occurrences is not aligned in the same way in the two data sets from our two annotators. The percentage of disagreement for articles reaches 40% for the definite article and about 27% for the indefinite article. Even though there are words with a higher percentage of disagreement, articles are reasonably important because of their frequency in the text.

The basic rule we used for articles was aligning them to the Czech head noun of corresponding grammatical construction, but in some cases it is not clear what is the head, and head selection depends on the particular specification of grammar and therefore, our annotators couldn't easily find the correct solution.

When we focused on identifying cases that cause troubles we noticed that one of the reasons for disagreement were situations where words are changing their POS during the translation process. For instance English idiomatic expressions such as *We learned a lesson...* are translated using a single verb in Czech (*Jsme se poučili...*, lit. We-have$_{past}$ ourselves$_{aux.refl.}$ taught$_{past}$...) or vice versa.

3.2. Verbs and Their Belongings

Verb tenses and usage of auxiliaries to express them is different in Czech and English. Although rules for aligning auxiliary verbs (such as *have*, *would*, *are*, etc.) were created thoughtfully, these words caused a lot of headache for our annotators. Just the word *be* made it to the top ten most frequently misaligned words, but the percentage of disagreement for other auxiliary verbs is high as well. (English *be* 63%, *is* 36%, *'s* 23%, *have* 49%, *are* 45%, *would* 39% and Czech *bude* 39%.)

Also alignment for the word *to*, which can indicate the infinitive form of a verb or it can be used as a preposition, was one of the less successful; every other appearance of this word wasn't aligned correctly.

Pronouns represent an area where our guidelines weren't detailed enough. Czech is a pro-drop language, pronouns representing the subject are usually left out but the morphology of the verb indicates explicitly which pronoun was meant. Our guidelines were not specific about alignment of the corresponding English pronouns (keep unaligned or align to the verb) so the inter-annotator disagreement reached highest numbers here (it 66%, he 94%, they 80%, It 82%, We 88%, He 91% etc.) A similar situation arose with *se* (64%), the Czech reflexive pronoun that has no real equivalent in English.

3.3. Punctuation

The last important issue worth mentioning is punctuation. When we take a closer look at the most common misaligned words, we realize that the most problematic token is the comma with the following percentage of disagreement: in English 9% and in Czech 20%. A part of this disagreement is caused by different decisions by our annotators on how to deal with this character in case where there is a comma in Czech but no in English. One of our annotators aligned commas to a particular conjunction whilst the other left them unconnected.

The type of text in the corpus (economic-related) highlights some specific problems of alignment. The Czech period character (.) was misaligned in 4% of cases and almost all of these cases where dates. Though in English the period is not used for writing the date, it is in Czech (*September 24* compared to *24. září*). This explains why this character is in the top ten of Czech misaligned words but not in English ones even though at first, the expectation would be approximately the same amount of periods in both data sets.

The *$* character (with its 12% rate of disagreement) is even more specific for the economic-related corpus since this corpus contains a lot of money related sentences. The use of the *$* character in English doesn't completely correspond to the use of the word *dolarů* in Czech texts since the word is usually used just once per sentence but the *$* character can be repeated in English several times. It is a challenge for our annotators to decide which pair they should connect together and whether to connect all *$* characters to one word or not.

	Intersection (1-1)			Union (n-n)		
	Prec	Rec	AER	Prec	Rec	AER
Baseline	97.4	57.6	27.4	65.9	86.7	25.5
Lemmas	97.9	75.0	15.0	77.1	89.8	17.2
Lemmas + Numbers	97.9	75.2	14.8	77.5	89.9	17.0
Lemmas + Singletons backed off with POS	97.4	75.8	14.6	77.8	88.5	17.4

Table 3: Improving GIZA++ alignments.

3.4. Other Tokens

Although Table 1 indicates a lot of disagreement caused by nouns, it is difficult to categorize them since usually these words didn't repeat too often and we dealt with rather specific issues.

When it comes to prepositions, often the difference between our two annotators was mainly made by a type of connection. Whereas one decided to link a preposition with a possible connection, the other used either a sure connection or didn't connect the word at all. When we compare a part of an English sentence and the same part of a corresponding Czech sentence *továren ve spojených státech* against English *U.S. factories* we can realize that some English language constructions don't require preposition, but the Czech language uses *v* or *ve*. In this particular example our annotators didn't agree on the type of connection which should be used. One of them marked the connection as possible while the other one as sure.

A comparable problem arises when indicating time frame. Where an English writer used just *1992* to indicate the year when something happened, translators enriched the Czech sentence with additional words *v roce 1992*. Both this and the previous example contain some additional words; the problem here is whether those additional words should be aligned and where.

Disagreement in alignment of conjunctions arises mainly due to a completely different sentence structure in Czech and English (even though the meaning is preserved). Thus, some conjunctions are changed or completely omitted and it is not quite clear if and where to align them.

4. Automatic Word Alignment

The state-of-the-art automatic word alignment systems are based on the GIZA++ toolkit (Och and Ney, 2003). We performed several experiments with corpus preprocessing to improve the quality of GIZA alignments.

4.1. Evaluation of Word Alignment

Table 3 summarizes the results of some of the techniques evaluated using the standard measures of precision, recall and alignment error rate (AER); see (Och and Ney, 2003). The measures evaluate GIZA-supplied alignments against manual "golden" annotations. Traditionally, golden annotations contain alignment points of two types only: POSSIBLE and SURE alignments. Precision errors penalize the algorithm for asserting an alignment that was not even possible while recall errors penalize the algorithm for omitting a sure alignment point. AER is a combination of precision and recall.

We created the golden alignments by combining our three types of connection from the two parallel annotations according the following rules: a connection is marked as sure if at least one of the annotators marked it as sure and the other also supported the link by any connection type. In all other cases (at least one annotator makes any type of link), the alignment is marked as possible. These rules slightly promote the introduction of sure alignments in the golden annotations, which is good, because too many possible alignments in the golden data weaken the metrics (there is no penalty for forgetting a possible connection).

GIZA++ toolkit is capable of guessing 1-n alignments (more target words get assigned to one source word). Typically, GIZA is used twice to obtain alignments in both directions. There are two common ways to obtain a joined alignment. Either the two directions are combined using intersection or using union. Intersection alignments have in general higher precision and lower recall compared to union alignments.

4.2. Improving GIZA++ Accuracy

Our experiments indicate that the key issue in automatic word alignment of Czech is the morphological richness of the language.

The baseline accuracy level (AER of 27% in Table 3) is achieved using input text that has been only tokenized. As documented in Table 4, lemmatization of the input text reduces the Czech vocabulary size to a half so that the vocabulary sizes of Czech and English become comparable. (The effect of lemmatization of the English is not that great.) The alignment task is thus greatly simplified and AER drops to about 15%. This matches with observations of (Popović et al., 2005) on Serbian-English machine translation task.

		Czech	English
	Sentences	21,141	
	Running Words	475,719	494,349
	Running Words without Punct.	404,523	439,304
Baseline	Vocabulary	57,085	30,770
	Singletons	31,458	14,637
Lemmas	Vocabulary	28,007	25,000
	Singletons	13,009	11,873
Lemmas	Vocabulary	15,041	13,150
+ Singletons	Singletons	12	2

Table 4: Characteristics of the Prague Czech-English Dependency Treebank 1.0.

Another great saving in vocabulary size can be achieved by replacing all words occurring only once (singletons) with a special symbol representing their part of speech. We still

		Baseline		Improved	
Humans	GIZA++	en	cs	en	cs
Problems	Problems	14.3	15.5	14.3	15.5
Problems	OK	0.1	0.1	0.2	0.1
OK	Problems	38.6	35.7	25.2	25.0
OK	OK	46.9	48.7	60.4	59.4

Table 5: Percentage of English (en) and Czech (cs) tokens where the alignment was difficult for humans and/or for GIZA++.

observe some improvement in alignment quality, but of a much smaller magnitude and using the intersection technique only.[1]

We also tried to use a common symbol for all numbers (provided that there is equal number of numbers in the Czech and the corresponding English sentence). This technique brings again a small improvement of AER, this time equally for intersection and union alignments.

Our results are fairly comparable to results achieved on other language pairs, if we use the lemmatization. (Och and Ney, 2003) report a detailed comparison of AER using various algorithms and various corpus sizes for German-English and French-English corpora. (Mihalcea and Pedersen, 2003) report on a shared task of many systems aiming at Romanian-English and French-English alignments. (Martin et al., 2005) is a similar report focused on languages with scarce resources.

4.3. Limits of Automatic Word Alignment

Table 5 analyzes the difficulty of the alignment of English and Czech tokens when aligned by humans (comparing the human alignments against each other) and when aligned by GIZA (comparing GIZA alignments to the golden annotation derived from both of the human alignments). As we see, around 15% of (Czech or English) tokens are difficult already for humans, since they do not agree on the alignment. (By chance, in a tiny portion of such tokens, GIZA finds the "correct", i.e. merged, alignment.)

Around 36–39% of tokens are easy for humans but GIZA misaligns them when the baseline method is employed. With the improved Lemmas+Singletons method, only 25% of tokens fall into this category. However, it should be emphasized that the improvement does never occur on tokens difficult for humans.

The results in Table 5 can be also put differently: 38% of the tokens where GIZA (using the improved method) fails to find the correct alignment are difficult already for humans.

A more detailed analysis of the tokens where GIZA had problems and humans aligned them in accordance revealed that the major contribution comes again from articles in English and commas in Czech.

We tried to tackle the problem with articles by removing them completely before running GIZA. The AER from this experiment was unfortunately worse than without circumventing articles (evaluated against golden annotations with removed articles, as well as evaluated against full golden annotations with articles aligned to the governing Czech noun using an independent rule). (Popović et al., 2005) report that removing articles helped in English to Serbian machine translation on a corpus of limited size. The positive effect vanished with corpus of about 2,500 sentences.

4.4. Summary

We are studying the problem of word alignment of the Czech and English languages. Inter-annotator agreement and observed difficult cases are examined in a closer detail. Results of the state-of-the-art automatic word alignment system GIZA++ are provided, including some important preprocessing tips to improve the accuracy.

An important observation comes from the analysis of errors of the automatic procedure and disagreement in human annotation. We document that nearly 38% of tokens where GIZA++ makes errors are difficult for humans already. This leads us to the conclusion that in order to achieve further improvements of automatic word alignment, a slight redefinition of the task is to be sought for. Otherwise, the rate of human disagreement would pose an unsurpassable boundary on the achievable accuracy.

5. Acknowledgement

The work on this project was supported by the grants Collegium Informaticum GACR 201/05/H014, GAUK 351/2005 and DAAD A/05/15151. Ondřej Bojar is also very grateful for the support of Prof. Hermann Ney's team during his stay at RWTH Aachen University.

6. References

Martin Čmejrek, Jan Cuř´ın, Jiř´ı Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of LREC 2004*, Lisbon, May 26–28.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 65–74, Ann Arbor, Michigan, June. Association for Computational Linguistics.

I. Dan Melamed. 1998. Annotation Style Guide for the Blinker Project. Technical Report IRCS-98-06, IRCS.

Rada Mihalcea and Ted Pedersen. 2003. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT/NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, May.

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

Maja Popovi´c, David Vilar, Hermann Ney, Slobodan Joviči´c, and Zoran Šari´c. 2005. Augmenting a small parallel text with morpho-syntactic language. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 41–48, Ann Arbor, Michigan, June. Association for Computational Linguistics.

[1] The drop in union AER using singletons is caused by a significant drop in recall. Using singletons helped GIZA to work similarly in both directions and thus produce a union alignment with about 300 points fewer in the evaluation data. Unfortunately some of those points were classified as sure in the golden data.