ПОВЫШЕНИЕ КАЧЕСТВА ВЫРАВНИВАНИЯ ПО ПРЕДЛОЖЕНИЯМ ДЛЯ ПАРАЛЛЕЛЬНОГО АНГЛО-РУССКОГО КОРПУСА ПРИ ПОМОЩИ ЧАСТЕРЕЧНОЙ РАЗМЕТКИ

INCREASING SENTENCE ALIGNMENT QUALITY IN PARALLEL ENGLISH-RUSSIAN CORPUS THROUGH THE USE OF PART-OF-SPEECH TAGGING*

Кутузов А.Б. (akutuzov72@gmail.com)
НИУ ВШЭ, Москва, Россия

Ключевые слова: параллельные корпусы, выравнивание по предложениям, частеречная разметка, расстояние Левенштейна

Kutuzov A.B. (akutuzov72@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

Keywords: parallel corpora, sentence alignment, part-of-speech tagging, Levenshtein distance

Abstract: The present paper introduces approach to improve English-Russian sentence alignment, based on POS-tagging of automatically aligned (by HunAlign) source and target texts. The initial hypothesis is tested on a corpus of bitexts. Sequences of POS tags for each sentence (exactly, nouns, adjectives, verbs and pronouns) are processed as “words” and Damerau-Levenshtein distance between them is computed. This distance is then normalized by the length of the target sentence and is used as a threshold between supposedly mis-aligned and “good” sentence pairs. The experimental results show precision 0.81 and recall 0.8, which allows the method to be used as additional data source in parallel corpora alignment. At the same time, this leaves space for further improvement.

Introduction

Parallel multilingual corpora have long ago become a valuable resource both for academic and for industrial computational linguistics. They are employed for solving problems of machine translation, for research in comparative language studies and many more.

One of difficult tasks in parallel multilingual corpora building is alignment of its elements with each other, that is establishing a set of links between words and phrases of source and target language segments (Tiedemann, 2003). Alignment can be done on the level of words, sentences, paragraphs or whole documents in text collection. Most widely used are word and sentence alignment, and the present paper deals with the latter one.

Word alignment is an essential part of statistical machine translation (SMT)

* The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2013.

workflow. However, usually it can only be done after sentence alignment is already present. Accordingly, there have been extensive research on the ways to improve it.

Basic algorithm of sentence alignment simply links sentences from source and target text in order of their appearance in the texts. E.g., sentence number 1 in the source corresponds to sentence number 1 in the target etc. But this scheme by design can't handle one-to-many, many-to-one and many-to-many links (a sentence translated by two sentences, two sentences translated by one, etc) and is sensitive to omissions in source or translated text.

Mainstream ways of coping with these problems and increasing alignment quality include considering sentence length (Gale and Church, 1991) and using bilingual dictionaries (Och and Ney, 2000) or cognates (Simard et al, 1993) to estimate the possibility of sentences being linked. Potemkin and Kedrova (2008) showed that these ways provide generally good results for Russian as well.

But often this is not enough. Sentence length can vary in translation, especially when translation language is typologically different from the source one. As for bilingual dictionaries, it is sometimes problematic to gather and compile a useful set of them.

Thus, various additional methods were proposed, among them using part-of speech data from both source and target texts. It is rather commonplace in word alignment (Tiedemann, 2003; Toutanova, 2003). Using part-of speech tagging to improve sentence alignment for Chinese-English parallel corpus is presented in (Chen and Chen, 1994). In the current paper we propose to use similar approach in aligning English-Russian translations.

Setting up the experiment

We test the part-of-speech based approach to improve quality of sentence alignment in our parallel corpus of learner translations available at http://rus-ltc.org. Only English to Russian translations were selected, as of now. The workflow was as follows.

All source and target texts were automatically aligned with the help of HunAlign software (Varga et al 2005) together with its wrapper LF Aligner by András Farkas (http://sourceforge.net/projects/aligner). The choice of aligner was based on high estimation by researchers (Kaalep and Veskis, 2007; Abdul-Rauf et al, 2010) and its open-source code.

HunAlign uses both bilingual dictionaries and Gale-Church sentence-length information. Its results are quite good, considering the noisiness of our material. However, about 30% of sentences are still mis-aligned. The reasons behind this are different, but mostly it is sentence splitter errors, omissions or number of sentences changing during translation. Here is a typical example:

(1) *“And these two fuels are superior to ethanol, Liao says, because they have a higher energy density, do not attract water, and are noncorrosive”.* ↔ *“Эти два вида топлива явно превосходят этанол по своим свойствам.”*

0 ↔ *“По словам Ляо, они обладают более высокой энергетической плотностью, не содержат воду, а значит некоррозийные.”*

The translator transformed one English sentence into two Russian sentences.

Consequently, aligner linked the first Russian sentence to the source one, and the second sentence is left without its source counterpart (null link). It should be said that in many cases HunAlign manages to cope with such problems, but not always, as we can see in the table above.

The cases of mis-alignment must be human corrected, which is very time-expensive, especially because there is no way to automatically assess the quality of alignment. HunAlign's internal measure of quality is often not very helpful. For example, for the first row of the table above it assigned rather high quality mark of 0.551299. Trying to predict alignment correctness with the help of Hun quality mark only for the whole our data set resulted in precision 0.727 and recall 0.548, which is much lower than our results presented below.

We hypothesize that source and target sentence should in most cases correspond in the number and order of content parts of speech (POS). This data can be used to trace mis-aligned sentences and perhaps to find correct equivalents for them. In order to test this hypothesis, our source and target texts were POS-tagged using Freeling 3.0 suite of language analyzers (Padro and Stanilovsky, 2012). Freeling gives comparatively good results in English and Russian POS-tagging, using Markov trigram scheme trained on large disambiguated corpus.

Freeling tag set for English follows that of Penn TreeBank, while Russian tag set corresponds to EAGLES recommendations for morphosyntactic annotation of corpora (http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html). It is not trivial to project one scheme onto another completely, except for the main content words – nouns, verbs and adjectives. Moreover, these three parts of speech are the ones used in the paper by Chen and Chen (1994), mentioned above. So, the decision was made to take into consideration only the aforementioned lexical classes, with optional inclusion of pronouns (in real translations they often replace nouns and vice versa).

Thus, each sentence was assigned a “POS watermark”, indicating number and order of content words in it. Cf. the following sentence:

(2) *“Imagine three happy people each win $1 million in the lottery.”*

and its “POS watermark”:

VANVNN,

where N is noun, A is adjective and V is verb.

Here is the same analysis for its Russian translation counterpart:

(3) *“Представим себе трех счастливых людей, которые выиграли в лотерею по миллиону долларов.”*

Corresponding “POS watermark”:

VPANVNNN,

where N is noun, V is verb, A is adjective and P is pronoun.

Nouns and verbs are marked identically in Penn and EAGLES schemes. Adjectives in Penn are marked as JJ, so this mark was corrected to A, which is also the mark for adjectives in EAGLES. We considered to be 'pronouns' (P) those words which are marked as “E” in EAGLES and “PRP” in Penn.

Thus, each content word is represented as one letter strictly corresponding to one lexical class. Therefore our “POS watermark” can be thought of as a kind of “word”. The

difference between these “words” is computed using Damerau-Levenshtein distance (Damerau, 1964). Basically, it is the number of corrections, deletions, additions and transpositions needed to transform one character sequence into another. We employ Python implementation of this algorithm by Michael Homer (published at http://mwh.geek.nz/2009/04/26/python-damerau-levenshtein-distance).

According to it, the distance between POS watermarks of two sentence above is 2. It means we need only two operations – adding one pronoun and one noun – to get target POS structure from source POS structure. At the same time, the distance between VPVNANNNNNNNNNVN and NVNNANANANN is as high as 10, which means that POS structures of these sentences are quite different. Indeed, the sentences which generated these structures are obviously mis-aligned:

(4) *“If a solar panel ran its extra energy into a vat of these bacteria, which could use the energy to create biofuel, then the biofuel effectively becomes a way to store solar energy that otherwise would have gone to waste.”* ↔ *“Однако они вырабатывают энергии больше, чем требуется.”*

One can suppose that there is correlation between Damerau-Levenshtein distance and the quality of alignment: the more is the distance the more is the possibility that the alignment of these two sentences has failed in one or the other way. In the following chapter we present the results of the preliminary experiment on our parallel texts.

The results

We performed testing of the hypothesis over 170 aligned English-Russian bi-texts containing 3263 sentence pairs. As of genres of original texts, they included essays, advertisements and informational passages from mass media. The dataset was hand-annotated and mis-aligned sentence pairs marked (663 pairs, 20% of total dataset).

Damerau-Levenshtein distances for all sentences were computed and we tried to find optimal distance threshold to cut “bad” sentence pairs from “good” ones. For this we used Weka software (Hall, 2009). The results were evaluated with 10-fold cross-validation over the entire dataset.

Initially, on the threshold 7 we achieved precision 0.78, recall 0.77 and F-measure 0.775 for the whole classifier. F-measure for detecting only mis-aligned sentences was as low as 0.464.

In order to increase the quality of detection we tried to change the settings: first, to change the number of “features”, i.e., parts of speech considered. “Minimalist” approach with only nouns and adjectives lowered F-measure to 0.742. However, considering nouns, adjectives and verbs without pronouns seemed more promising: using the same distance threshold 7 we got precision 0.787 and recall 0.78 with F-measure 0.783. F-measure for detecting mis-aligned sentences also got slightly higher, up to 0.479. So, general estimate is even higher than when using pronouns.

Moving further in an effort to improve the algorithm, we found that Damerau-Levenshtein distance shows some kind of dis-balance when comparing short and long “words”. Short “words” receive low distance estimates simply because the number of characters is small and it's “easier” to transform one into another, even if the

“words” are rather different. At the same time, long “words” tend to receive higher distance estimates because of higher probability of some variance in them, even if the “words” represent legitimate sentence pairs. Cf. the following pairs:

- distance between PVPVAA and ANAN is estimated as 5,
- distance between NNNNVAANNVVNNVNNNVV and NNNNVANANPANNANVN is estimated as 7.

Meanwhile, the first sentence pair is in fact mis-aligned, and the second one is quite legitimate. It is obvious that “word” length influences results of distance estimation and it should be somehow compensated.

Thus, the penalty was assigned to all distances, depending on the length of original sentences. Then this “normalized” distance was used as a threshold. We tried employing the length of the source sentence, of target sentence and the average of both. The length of the target (translated) sentence gave the best results.

So, the equation is as follows:

$$DLnorm = \frac{DL(sP, tP)}{LEN(tP)},$$

where DLnorm is “normalized” distance, DL is original Damerau-Levenshtein distance, sP is “POS watermark” for source sentence, tP is “POS watermark” for target sentence and LEN is length in characters.

With nouns, verbs, adjectives and pronouns this normalization gives considerably better results:

Precision 0.813

Recall 0.802

F-Measure 0.807

After removing pronouns from consideration, at the optimal threshold of 0.21236, recall gets slightly higher:

Precision 0.813

Recall 0.803

F-Measure 0.808

Even “minimalist” nouns-and-adjectives approach improves after normalization:

Precision: 0.792

Recall: 0.798

F-Measure: 0.795

Overall results are presented in the table 1.

Method	Precision	Recall	F-measure
Nouns, adjectives, verbs and pronouns without length penalty	0.78	0.77	0.775
Nouns, adjectives and verbs without length	0.787	0.78	0.783

penalty			
Nouns and adjectives without length penalty	0.764	0.728	0.742
Nouns, adjectives, verbs and pronouns with target length penalty	0.813	0.802	0.807
Nouns, adjectives and verbs with target length penalty	0.813	0.803	0.808
Nouns and adjectives with target length penalty	0.792	0.798	0.795

Table 1. Overall performance of pairs classifier depending on the method

Methods without target length penalty provide considerably lower overall performance, thus, methods with the penalty should be used.

Depending on particular aim, one can vary the threshold used in classification. In most cases, mis-aligned pairs are of more interest than “good pairs”. If one's aim is to improve precision of “bad pairs” detection, the threshold of 0.8768 will give 0.851 precision for this, at the expense of recall as low as 0.1. If one wants more balanced output, the already mentioned threshold of 0.21236 is optimal, providing mis-aligned pairs detection precision of 0.513 and recall of 0.584.

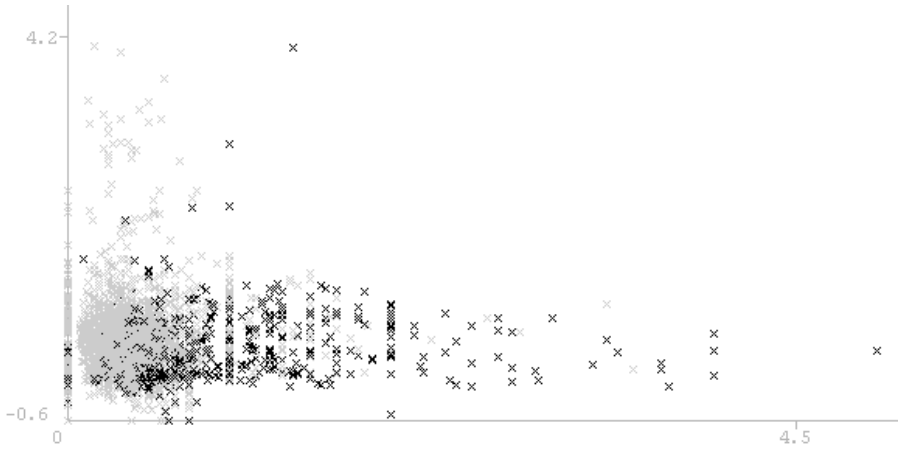

Fig. 1. Levenshtein distance (X axis) and alignment correctness (color)

Figure 1 presents distribution of “good” and “bad” pairs in our data set in relation to Damerau-Levenshtein distance (X axis). Correctly aligned pairs are colored gray and incorrectly aligned ones black. Correlation between alignment correctness and Levenshtein value can be clearly seen. At the same time, internal HunAlign quality measure (Y axis) does not show any stable influence on alignment correctness, as we already mentioned above.

Discussion and further research

Number and order of POS in source and target sentences in English-Russian translations do correspond in most cases. The method of checking Damerau-Levenshtein distance between POS “watermarks” of source and target sentences can be applied for detecting mis-aligned sentence pairs as an additional factor, influencing the decision to

mark the pair as “bad”.

However, some pairs show anomalies in this aspect. For example, the pair below is characterized by normalized POS Damerau-Levenshtein distance of enormous 2.6, however, human assessor marked it as “good”:

(5) *“An opinion poll released by the independent Levada research group found that only 6% of Russians polled sympathised with the women and 51% felt either indifference, irritation or hostility.”* ↔ *“А вот 51% опрошенных испытывают к ним равнодушие и даже враждебность.”*

Translator omitted some information she considered irrelevant, but the pair itself is aligned correctly.

On the other hand, cf. two consecutive pairs below:

(6) *“The British Museum? The Louvre?”* ↔ *“Британский музей?”*

“The Metropolitan?” ↔ *“Лувра?”*

Normalized distance for the first pair is 0.3333, and this correctly classifies it as “bad”. The second target sentence must have belonged to the first pair and the second pair is obviously bad, but its distance equals to zero (because both part contain exactly one noun), so it will be incorrectly classified as “good” with any threshold.

Such cases are not detected with the method described in this paper.

Our plans include upgrading this method from one passively marking mis-aligned pairs and leaving the actual correction to human to the one actively searching for possible equivalent candidates among other sentence pairs, especially among those with null links. The difficult part here is designing the method to deal with “partially correct” alignment, for example, like in the pair below:

(7) *“The magic number that defines this “comfortable standard” varies across individuals and countries, but in the United States, it seems to fall somewhere around $75,000.”* ↔ *“Волшебная цифра, которой определяется уровень комфорта, зависит от самого человека, а также от страны, в которой он проживает.”*

In the experiment above we considered such pairs to be mis-aligned. But ideally, the second part of the source sentence should be detached and start “looking for” appropriate equivalent. Whether this can be done with the help of POS-tagging (or, perhaps, syntactic parsing), further research will show.

The same is true about the possibility to apply this method to Russian-English translations or translations between typologically distant languages.

Conclusion

In this paper, approach to improve English-Russian sentence alignment was introduced, based on part-of-speech tagging of automatically aligned source and target

texts. Sequences of POS-marks for each sentence (exactly, nouns, adjectives, verbs and pronouns) are processed as “words” and Damerau-Levenshtein distance between them is computed. This distance is then normalized by the length of the target sentence and is used as a threshold between supposedly mis-aligned and “good” sentence pairs.

The experimental results show precision 0.81 and recall 0.8 for this method. This performance alone allows the method to be used in parallel corpora alignment, but at the same time leaves space for further improvement.

Reference

1. Kuang-hua Chen, Hsin-Hsi Chen. 1994. *A part-of-speech-based alignment algorithm.* Proceedings of the 15th conference on Computational linguistics, Kyoto, Japan.
2. Fred J. Damerau. 1964. *A technique for computer detection and correction of spelling errors.* Commun. ACM 7, 3, 171-176.
3. W.A. Gale and K.W. Church. 1991. *A program for aligning sentences in bilingual corpora.* In Meeting of the Association for Computational Linguistics, pages 177–184.
4. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten. 2009. *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, Volume 11, Issue 1.
5. Kaalep, Heiki-Jaan and Kaarel Veskis. 2007. *Comparing parallel corpora and evaluating their quality.* In Proceedings of MT Summit XI, pages 275–279, Copenhagen, Denmark.
6. G.E. Kedrova and S.B. Potemkin. 2008. *Alignment of un-annotated parallel corpora.* In Papers from the annual international conference 'Dialogue', issue 7 (14). Moscow: 431-436.
7. P. Lambert, S. Abdul-Rauf, M. Fishel., S. Noubours, R. Sennrich. 2010. *Evaluation of sentence alignment systems.* Fifth MT Marathon. Le Mans, France.
8. Franz J. Och and Hermann Ney. 2000. *A comparison of alignment models for statistical machine translation*. In COLING ’00: The 18th International Conference on Computational Linguistics, pages 1086–1090, Saarbrucken, Germany, August
9. Lluís Padró and Evgeny Stanilovsky. 2012. *FreeLing 3.0: Towards Wider Multilinguality.* Proceedings of the Language Resources and Evaluation Conference (LREC-2012) ELRA. Istanbul, Turkey.
10. Michel Simard, George F. Foster, Pierre Isabelle. 1992. *Using Cognates to Align Sentences in Bilingual Corpora.* In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.
11. Kristina Toutanova, H. Tolga Ilhan, and Christopher D.Manning. 2003. *Extensions to HMM-based statistical word alignment models.* In Proceedings of Empirical Methods in Natural Language Processing, Philadelphia, PA.
12. D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy. 2005. *Parallel corpora for medium density languages.* In Proceedings of the RANLP: 590-596.