Natural Language Processing for the Translation Class

Lars Ahrenberg[1] *and Ljuba Tarvi*[2]

(1) Department of Computer and Information Science, Linköping University
(2) Helsinki University

lars.ahrenberg@liu.se, ljuba.tarvi@welho.com

ABSTRACT

We propose a system for use in translation teaching with automatic support for alignment and comparative assessment of different translations. A primary use of this system is for discussion in class and comparison of student translations from a given source text, but it may also be used to study and compare differences between published translations. We describe the intended functions of the system and give suggestions on its design and architecture. We also discuss the degree of automation that can be expected and report results from a small indicative study focused on word alignment performance.

KEYWORDS: natural language processing, translation teaching, translation assessment, alignment

Lars Ahrenberg and Ljuba Tarvi 2013. Natural language processing for the translation class. *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013.* NEALT Proceedings Series 17 / Linköping Electronic Conference Proceedings 86: 1–10.

1 Introduction

With the advent of computational aids for translators, such as translation memory systems, terminology management systems and corpus search tools, the need to teach the use of such tools in translator training has also been recognized. From the perspective of natural-language processing, however, these tools are not very sophisticated. In particular, the technologies that have propelled the fast developments in machine translation have not been used as much as they could be.

In this paper we give a proposal for a system that can support the assessment and classroom discussion of students' translations. Crucial to both aims is having the translations aligned at the word and segment levels with the source text. From this alignment global metrics of the student translations can be computed, helping them to understand the style of their translation in relation to other translations, including published ones. By using quantitative measures that have been shown to correlate well with qualitative judgements, it also helps the grading of students' translations. Moreover, the alignment can support different kinds of visualizations of the students' work. Our hypothesis is that a sufficient alignment quality can be obtained by using a combination of automatic and interactive methods.

In the following, we first, in section 2, report on related work and then proceed, in section 3, to give an overview of the design and functions of the proposed system. In section 4 we describe preliminary results of a small experiment on alignment of texts that have been, or could be used for class instruction. Section 5, finally, holds our conclusions.

2 Related work

Translation memory systems and other CAT (Computer-Aided Translation) tools are increasingly being used in translator training, and the creation and use of corpora has been a common interest for translation studies, translator training and computational linguistics for several years (e.g. Zanettin et al., 2003). In translator training the corpora are mostly seen as resources for the student to use when translating (Lopez-Rodriguez and Tercedor-Sanchez, 2008; Pastor and Alcina, 2009). Proposals have also been made to use e-learning environments for specialized translation courses, where students' translations can be collected and compared by all course participants (Fictumová, 2007).

Also in our proposal immediate comparisons and assessments of students' translations are offered as a class-based activity. A system with some similarity is reported in Shei and Pain (2002: 323) who describe an "intelligent tutoring system designed to help student translators learn to appreciate the distinction between literal and liberal translation". Their system allows students to compare their own translations with reference translations and have them classified in terms of categories such as literal, semantic, and communicative. The comparisons are made one sentence at a time, using the Dice coefficient, i.e., by treating the sentences as bags of words. To contrast, our proposal uses more advanced computational linguistics tools, offers more teacher involvement, and provides text level assessment based on token alignments.

Our proposal relies heavily on recent advances within computational linguistics. In particular, it can be viewed as a test bed for current alignment technology. In addition it draws on the Token Equivalence Method (TEM; Tarvi, 2004) where the idea that translation correspondence at the token level is useful for the characterization and assessment of translations is developed (see section 3).

Alignment technology has advanced considerably over the years and is still an active area of research (Tiedemann, 2011). Word alignment is an input technology for a range of bilingual or multilingual NLP tasks. Most prominent of these is perhaps statistical machine translation (SMT; Koehn, 2009) but also many others such as terminology extraction, lexicon generation, and the creation of parallel corpora and treebanks. As far as we are aware, however, there is no published work reporting on alignment technology for use in the translation class.

3 System overview

The proposed environment has a central system for the teacher and a number of client systems for the students. The students' systems can be designed somewhat similar to a translation memory, allowing alternative views of the source and target texts and, if need be, enforcing alignment of a student's work with the source text at an appropriate segment level (sentence or paragraph). When a student has finished a translation task, she will save her translation in an XML-based exchange format such as an XLIFF extension and make it available for the teacher, say, by uploading it through a web interface. For the rest of this section we will focus on the teacher system.

The teacher's system is equipped with several modules for text analysis, including tokenization, lemmatization, part-of-speech tagging, sentence and word alignment, where automatic tools are integrated into an interactive environment. Any output from an automatic component can be reviewed and changed by the teacher. The teacher's system also has components for visualization and joint display of the student translations.

When a text has been selected for a translation exercise, it will be segmented, tokenized and indexed. The teacher can prepare the system dictionary for the new text as required and identify multiword units, including idioms, as units of special interest. When translations are returned, the teacher is acting as a post-editing human agent who can combine both manners of assessment, computer-assisted and manual. After tokenization and indexing the translations can be analyzed in the same way as the source text and be aligned with it. The teacher reviews the alignments and corrects the errors.

Sentence alignment can be enforced for a given translation task, but if the teacher does not want that, sentence aligners usually perform well enough on the kind of short texts that are suitable for a translation class. Word alignment is a different matter. While error rates as low as 5% or less have been reported on some data sets (Liang et al. 2006; Moore et al. 2006), such figures are hard to achieve. Only practice can show what level of accuracy is actually required for the system to be useful and requirements may be different for classroom display and for grading purposes.

Alignment of a source text with several translations runs the risk that different translations segment and order the content in different ways so that no single segmentation of the source text can be taken as adequate for all translations. Within a text we can recognize segments, phrases, and tokens. Segments should be big enough to have one-to-one corresponding segments in all translations. Tokens are the smallest text units and phrases are made up of one or more tokens within a segment.

The source text is maintained as a single file. It is connected to the translations via alignment files, one for each translation. Alignments at both segment and token level are represented in the alignment files. Translations of a source phrase can be computed for each translation based on the token alignments. It may of course happen that some part of a selected phrase has not been translated, or that the alignment contains more tokens than necessary. This information can be collected during the retrieval process and be displayed with the retrieved phrase.

3.1 Translation views

We imagine the system to support different views of the translations. A basic view is the **segment view** where a segment from the source text is displayed with one or more corresponding segments from the translations. This is the easiest one to implement as it only requires a correct segment alignment, where a segment may be a sentence, or a short paragraph. Words and phrases of interest in a source segment can be high-lighted, but the corresponding translations have to be recognized by the students without help from the system.

Another view is the **token view**, where a word or phrase at a specific position in the source text is singled out and its different translations are displayed. The display of translations can be restricted to an arbitrary subset of the translations, and the context can also be varied, say, to one or more segments or in terms of bytes. The matching tokens can be high-lighted against the still visible context.

A **type view** of the data is of interest when some word or phrase is used in different parts of the source text. Apart from just listing the different translations and their distribution on the students' texts, frequency tables are also compiled.

In addition, the system can display the outcomes of the different metrics that are described in the following section. These offer a **global view** of the translations, such as the amount of information from the source that are kept in the different translations. Such data can be displayed as a table, like Table 1 below in section 3.3.

3.2 Assessment and grading

There are a number of global metrics that can be computed from a word alignment. Here we follow the TEM framework. In Tarvi (2004) the TEM was used for comparing the classical Russian novel in verse by A. Pushkin *Eugene Onegin* (1837) and its then existing English translations. The quantitative figures calculated on 10% of the text of the novel showed a very good fit with the results obtained elsewhere on the same material by conventional comparative methods. Also, it could answer the question of which one of all the translations is the closest to the original, in both content and form.

Methodologically, the TEM focuses on what has been kept in translation. Two basic analytical planes are considered – content and formal. The lexical content of the original retained in its translation(s) is calculated as a percentage of the former. Several means of comparative assessment, in TEM referred to as 'frames', can be used, with the cumulative result – Translation Quotient (TQ) – calculated as an arithmetic mean of the percentages in all frames. There are also optional frames that focus on other characteristics of the translations that reflect the translator's style. In some analytical frames, the results are calculated as absolute numbers.

To illustrate the method, an eight-word excerpt (One LIX: 1-2) and the following five translations of *Eugene Onegin* are used: the translation by Vladimir Nabokov (1964), and the four latest versions – by Tom Beck (2004), Stanley Mitchell (2008), Henry M. Hoyt (2008), and D.M. Thomas (2011). The source sentence contains three Subject (S) – Predicate (P) groups, one Conjunction (C), and one Attribute (A):

Pushkin:	*1:Proshla*	*2:lyubov,*	*3:yavilas'*	*4:muza,*	*5:i*	*6:projasnilsya*	*7:tyomnyi*	*8:um.*
	[passed]	[love]	[appeared]	[muse]	[and]	[cleared up]	[dark]	[mind]
	P1	S1	P2	S2	C	P3	A	S3

The translations are shown with the alignments in the direction from translation to source inserted (punctuation marks are ignored). Thus, the first link (1-2) associated with Nabokov's translation says that the first word in the translation corresponds to the second word of the original. A zero (0) indicates that a word has no correspondent. For clarity multiword translations have been underlined and tokens with null links are indicated in bold:

Nabokov: Love passed, **the** Muse appeared, and **the** dark mind cleared up.

1-2 2-1 3-0 4-4 5-3 6-5 7-0 8-7 9-8 10-6 11-6

Beck: **Once** love **had** passed, **the** muse **then** surfaced, **the** darkness **in my** mind **had** cleared.

1-0 2-2 3-0 4-1 5-0 6-4 7-0 8-3 9-0 10-7 11-0 12-0 13-8 14-0 15-6

Hoyt: Love past, **the** muse **has** made appearance, and **the** dark mind **has** changed to light;

1-2 2-1 3-0 4-4 5-0 6-3 7-3 8-5 9-0 10-7 11-8 12-0 13-6 14-6 15-6

Mitchell: Love passed, **the** Muse resumed dominion and cleared **the** darkness **from my** mind,

1-2 2-1 3-0 4-4 5-3 6-3 7-5 8-6 9-0 10-7 11-0 12-0 13-8

Thomas: Love **as she** leaves lets in **the** Muse, and clarity **once more I find**.

1-2 2-0 3-0 4-1 5-3 6-3 7-0 8-4 9-5 10-6 11-0 12-0 13-0 14-0

Note the mode of alignment suggested here: only the meaningful denotative tokens are aligned, while added grammar tokens, such as *had* or *the,* are given null alignments. Thus, although token *6:projasnilsya* has been rendered as *cleared up* (Nabokov), *had cleared* (Beck), *changed to light* (Hoyt), *cleared* (Mitchell), and even as *clarity* (Thomas), all these renderings are viewed as retaining the denotative meaning of the original token. The connotative shades of meaning most suitable for the outlined goals can be discussed in class.

When employed manually, TEM employs such operations as consecutive numbering of the tokens in the source text; finding correspondences between the source and target tokens, identifying grammar tokens, parts of speech and syntactic positions, and calculating the obtained results as counts, percentages and Translation Quotients (TQ) for the purpose of grading. Therefore, the method generates absolute score (overall estimates) based on relative scores in separate frames (see Table 1).

All of this work can be automated, promising a substantial reduction in the time to perform a TEM analysis. Some of the automatic modules, given the current state-of-the-art will introduce a high number of errors, however, and for this reason, their output needs to be reviewed and corrected. The most critical one is the word alignment.

3.3 TEM Frames

In automatic mode, the (corrected) alignment files are used to calculate how much of the original information has been retained in the translations. Two content frames are used here – one basic, and one optional. The **basic content frame** (BCF) computes the number of source tokens that are part of a non-null alignment. This figure is then rendered as a percentage of the number of content tokens in the original. As is seen Nabokov, Hoyt and Mitchell translated all eight tokens and, hence, scored 100% each, Beck ignored *5:i* (87%), while Thomas has left out *7:tyomnyi 8:um* (75%).

The **optional content frame** (OCF) is a useful tool in additional assessment as it shows what has been added to the translation or that have no counterparts in the source texts. This can be calculated as an absolute number. Nabokov and Hoyt added no excessive content tokens, Mitchell added one (*from*), Beck – three, (*once, then, in*) Thomas – six (*as, she, once, more, I, find*). Note that not all null-aligned tokens are relevant to the OCF; grammar tokens that are required or suggested by the target language grammar are not counted. Thus, the OCF as other formal frames require an explicit recognizer for these tokens.

The formal frames pertains to the formal aspects of the translations in comparison with the original. In this analysis, there is a basic frame and two optional ones.

The **basic formal frame** (BFF), has the grammar tokens, – articles, tense markers, etc. at the centre of attention. Also these (*the, had, has, my*) can be seen to be employed in different quantities in the translations above: Thomas used only one, Nabokov – two, Mitchell – three, Hoyt – four, Beck – five. This frame, like other obliquely source-dependent frames, can say something about the translator's (or student's) individual style.

The **optional formal frame I** (OFF1) monitors another aspect of a translation. It counts the content tokens that are rendered with the same part of speech (PoS) in the translation as in the source. It is expressed as a percentage of all content tokens of the source. It is to be noted, that, like in other optional frames, the results reflect the translator's strategies to render the original rather than the intrinsic qualities of latter. Nabokov used in all eight tokens the same part of speech as in the original, Hoyt in seven (he used a participle, *past*, instead of the verb for 1:*Proshla*, Beck and Mitchell rendered the adjective *7:tyomnyi (dark)* with a noun *darkness*, while Thomas kept the PoS for only the first five tokens of the original.

Another way of gauging the 'presence' of the original in its translation is to register the syntactic changes. It is indisputable that there are certain syntactic changes in translations that are inevitable, due to the grammatical requirements of the target language, like, for instance, source tokens 1-2, 3-4, and 6-8 here, which can be translated into English only in a reverse order. However, translators have the option to reformulate and go beyond what is minimally required in rendering the contents of the source text.

If two tokens are rendered in the same sequence as in the original and preserve the same syntactic functions, they are considered kept. The **optional formal frame II** (OFF2) counts the number of such pairs and renders it as a percentage of all pairs. As could be expected, the most dramatic changes happened in the last group of tokens, Sts 6-7-8, with, for instance, St 8, originally Subject 3, rendered as Prepositional Objects (PO) by Beck and Mitchell; or St 7, originally an Attribute (A), rendered as a Direct Object (DO) by the same authors. Only Nabokov and Hoyt managed to have kept the attribute *dark* (St 7) in its original syntactic function.

To compute OFF2 automatically requires a good parser. Simpler measures that register reorderings from the alignments have been proposed in the literature, e.g. Kendall's tau or the LRscore (Birch and Osborne, 2011). These measures, while not using syntactic functions can still rank different translations with respect to the amount of reordering.

	BCF	OFF1	OFF2	**TQ**	OCF	BFF	**Rank**
Translator	%	%	%	**pp**	count	count	**N**
Nabokov	100	100	25	**75**	0	2	**1**
Beck	87	75	0	**54**	3	5	**4**
Hoyt	100	87	25	**70**	0	4	**2**
Mitchell	100	87	0	**62**	1	3	**3**
Thomas	75	62	0	**45**	2	1	**5**

Table 1. The TEM applied to eight words; assessment and grading.

3.4 Grading

Grading can be based on the frames. The TEM employs a measure called the **translation quotient** (TQ) which is calculated as the arithmetic mean of the percentages obtained in the frames. Moreover, as all translations can be given a rank for each frame, they can also be ranked from the TQ (see Table 1). After class discussion the students can revise their translations and one more monitoring can be carried out. The final grade, which can be an arithmetic mean of the home and class grades, is not only displayed but is registered automatically. If, at the end of class, the final grades are exhibited on screen in their ranking order, it is the best possible motivation for students to work diligently both at home and in class.

4 What word alignment technology to use?

As word alignment is crucial to the proposal, it is of interest to know what performance we can expect from currently available alignment systems, and what work is required from the teacher in order to get data of sufficient quality.

Word alignment systems usually give two kinds of output, token-oriented and type-oriented. The token-oriented alignment connects positions of the parallel texts, while the type-oriented output provides associations of words and phrases from the corpus as a whole, with or without probabilities. In the case of machine translation and the extraction of lexical data, the token-based alignment is not of primary interest; it is rather the word and phrase associations that can be derived from it. In our application both are relevant, but the token-based alignment is primary.

The most widely used word alignment systems, such as Giza++ (Och and Ney 2003) and its relatives, are statistical, learning word translation probabilities from parallel data. The alignment problem that we wish to find a solution for has the following characteristics:

- The source text is usually short, maybe in the range of 500–2000 words
- There are several translations and the parallel corpus to be aligned can be built from all the different translations and repeated versions of the source text
- Source and target languages are known so available resources in the form of dictionaries, SMT phrase tables, morphological analyzers, taggers, named entity recognizers, and parsers can be used

The fact that the texts are short speaks against using a statistical aligner. On the other hand, since the number of different translations can be high, data may still be sufficient for the exploitation of statistical tendencies. Also, we may augment the corpus with relevant portions of free parallel corpora, such as Europarl, based on lexical overlap. As the languages and source text are known in advance, word aligners that are based on generic resources such as dictionaries, syntactic pattern correspondences, and distortion distributions, the latter computed, say, from parallel treebanks, can also be employed. A framework using such resources is the "pressure aligner" of Esplà-Gomis et al. (2012).

We have made initial studies of alignment performance of Giza++ and a pressure aligner for two data sets. The purpose of these experiments is to find out what level can be reached with these systems. In particular we want to study the effect of the number of translations available for the statistical aligner, the effect of text-specific dictionaries for the pressure aligner, and the possibility to combine the two methods.

The first data set we have used is Russian–English; it comprises 17 stanzas (1085 tokens) from Eugene Onegin and eight different translations. We applied Giza++ (model 4) with standard settings to this data varying the size of the training corpus from one translation to all eight. As expected, performance improved with the addition of more translations; for one translation precision and recall are close to 30%, for eight translations it rises to 48%. We have not yet applied a pressure aligner to this data.

The second data set is English–Swedish with student translations from a translation class. The source text is made up of two short text snippets used in translation exercises, altogether 1234 tokens. There are three student translations and one published translation. To augment the corpus, two translations made by Google Translate and Microsoft Translator were added. The test set is the first short text with 452 tokens. We report precision and recall figures for six different set-ups in Table 2. The first two rows shows that more training data helps performance of the statistical system Giza++ (model

4). PA-1 is a pressure aligner with a dictionary for the most common English words and a short list of syntactic pattern correspondences. PA-2 has an added lexicon with words from the source text including correspondences found in the test set. The table also shows performance for the union and intersection of the two best aligners.

Id	System	Corpus size	Null links included		Null links excluded	
			Precision	**Recall**	**Precision**	**Recall**
1	Giza++[1 trl]	452	0.68	0.68	0.75	0.65
2	Giza++[5 trl]	2260	0.75	0.74	0.82	0.70
3	PA-1	452	0.50	0.55	0.82	0.49
4	PA-2	452	0.61	0.66	0.89	0.61
5	Union(2,4)	N.A.	0.74	**0.79**	0.78	**0.78**
6	Intersection(2,4)	N.A.	**0.87**	0.54	**0.98**	0.53

Table 2. Word alignment results for different systems. Best values are shown in bold.

5 Conclusions

We have presented an innovative concept for computer-aided translation teaching, based on existing token-based analyses of translations from computational linguistics and translation studies. As word alignment is the most crucial process for the proposal, we have also reported a pilot study on the feasibility of current alignment technologies for use in the system.

While the word alignment evaluation is small-scale, we believe it shows promising results. The statistical aligner improves when more translations are used, and the pressure aligner is able to take advantage of small increments to its dictionaries. In addition, they both find correspondences that the other aligner does not, so results can be further improved by combining them. With these small amounts of data, however, both aligners produce too many null links. That is why performance is better when only non-null links are considered. For post-editing, it is probably better to leave null links out, but for the test corpus at hand, this still means that at least some 200 links need to be added for a complete alignment. This is quite a lot of work, in particular if we consider that it should be multiplied by the number of translations at hand.

Still, we have not exhausted the potential of our word aligners. Performance is likely to improve by extending training data with open parallel resources for the statistical aligner, and using a much larger dictionary and phrase list for the pressure aligner. Also, as interactive word alignment can arguably be said to have some pedagogical value for the analysis of translations, this is work that may sometimes be performed by the students as a class-based activity.

References

Beck, Tom (2004) *Eugene Onegin by Alexander Pushkin*, Sawtry: Dedalus.

Birch, Alexandra and Osborne, Miles (2011) Reordering Metrics for MT. In Proceedings of the 49th Annual Meeting of the ACL, Portland, Oregon, USA.

Esplà-Gomis, Miquel, Sánchez-Martínez, Felipe, and Mikel L. Forcada (2012) Using external sources of bilingual information for in-th-fly word alignment. arXiv:1212.1192v2 [cs.CL] 7 Dec 2012, [last visited on 17/12/2012].

Fictumová, Jarmila (2007). Technology-enhanced Translator Training (Possible Pitfalls and Problems – A Case Study). Eldům, eldům - databáze publikovaných článků, 2007, Neuveden, Neuveden, pp. 1-12.

Hoyt, Henry M. (2008) *Alexander Pushkin. Eugene Onegin. A Novel in Verse,* Indianapolis: Dog Ear Publishing.

Koehn, Philipp (2010). *Statistical Machine Translation*, Cambridge University Press.

Liang, Percy, Taskar, Ben, and Klein, Dan (2006). "Alignment by Agreement." In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2006, 104-111.

López-Rodríguez, Clara Inés and Tercedor-Sánchez, María Isabel (2008) "Corpora and Students' Autonomy in Scientific and Technical Translation training" *Journal of Specialized Translation* (JoSTrans), Issue 09 (2008) pp. 2-19.

Mitchell, Stanley (2008) *Alexander Pushkin. Eugene Onegin*, London: Penguin.

Moore, Robert C, Yih, Wen-tau, and Bode, Anders (2006). "Improved Discriminative Bilingual Word Alignment." In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 2006, 513-520.

Nabokov, Vladimir (1975 [1964]). Nabokov, Vladimir. Eugene Onegin. Revised Edition. A Novel in Verse by Alexander Pushkin translated from the Russian, with a Commentary, by Vladimir Nabokov. In Four Volumes, New York: Boulinger Foundation.

Och, Franz Josef and Ney, Hermann (2003) A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics 29(1), 19-51.

Pastor, Verónica, and Alcina, Ampero (2009) Search techniques in corpora for the training of translators. *Proceedings of the workshop on NLP Methods and Corpora in Translation, Lexicography and Language Learning*, Borovets, Bulgaria, pp. 13-20.

Shei, Chi-Chiang and Pain, Helen (2002) "Computer-Assisted Teaching of Translation Methods." *Literary & Linguistic Computing*, Vol, 17, No 3 (2002), pp. 323-343.

Tarvi, Ljuba (2004) *Comparative Translation Assessment: Quantifying Quality*. Helsinki: Helsinki University Press.

Thomas, D.M. (2011) *Alexander Pushkin. Onegin*, London: Francis Boutle Publishers.

Tiedemann, Jörg (2011) *Bitext Alignment*. Morgan & Claypool Publishers.

Zanettin, Federico, Bernardini, Silvia, and Stewart, Dominic (Eds.) (2003). *Corpora in Translator Education*. Manchester: St Jerome Publishing.