Compiling a Turkish-English Bilingual Corpus and Developing an Algorithm for Sentence Alignment

Şerafettin Taşçı, A. Mustafa Güngör, Tunga Güngör
Boğaziçi University, Computer Engineering Dept., Istanbul, Turkey

Abstract: *In this paper, we discuss the compilation of a bilingual Turkish-English corpus and propose a method for sentence alignment based on location and length information in the texts. The content of the corpus was collected from several sources of different genre and it contains about 5 million words. To the best of our knowledge, this is the first comprehensive bilingual corpus between these languages. The proposed sentence alignment algorithm was tested on the corpus and success rates up to 96% were obtained.*

Keywords: *Sentence alignment, Bilingual corpus, Machine Translation*

1. INTRODUCTION

Sentence alignment is the task of determining sentence correspondences in a bilingual corpus and has a crucial role in corpus-based machine translation. Sentence alignment should be performed before the more ambitious task of word alignment. Basically, alignment aims to help the task of extracting structural information and statistical parameters from bilingual corpora.

The alignment process has some important challenges which make it difficult: First of all, most of the time sentences do not align 1-1. A sentence may be translated as 2-3 sentences in the other language, some part of a text may be deleted, or some additional sentences may be added to the text which has no match in the corresponding text. Even the existence of a small number of such sentences results in high deviations in the matching of sentence beads. Secondly, there is the problem of robustness. In real life, most of the texts have inconsistencies with their translations, such as the layout of texts, format differences, omission of some part of text and crossovers or inversions in text. The sentence alignment algorithms must be devised in such a way to deal with such diverse situations and problems. Finally, the problem of accuracy always exists. It is not easy to achieve perfect 100% accurate alignments even if the texts are clean and easy. Also the accuracies vary largely according to the input text. For example, an alignment algorithm may give successful results when applied on a scientific text, but its success may decline dramatically when applied on a novel or a philosophy text.

The aim of the research in this paper is two-fold: First, we aim at compiling a reliable and comprehensive bilingual Turkish-English corpus. This is the first step in obtaining an aligned parallel corpus via some alignment algorithms, which can later be utilized in developing corpus-based machine translation systems between these two languages. The second direction in this research is developing a sentence alignment algorithm for aligning Turkish and English texts and testing its applicability on the corpus. The motivation behind this research comes from the lack of studies related to Turkish language. Turkish belongs to the group of agglutinative languages where the affixation process is highly productive and also it can be characterized as a free word order language. It is desirable to take these characteristics into account while developing natural language processing systems. To the best of our knowledge, the corpus formed in this research is the most comprehensive Turkish-English bilingual corpus.

2. RELATED WORK

There are basically three approaches in sentence alignment algorithms. In length-based algorithms, the content of the text in terms of semantics is not taken into account. These algorithms make use of statistical methods and consider only the length of the sentences. Despite their simplicity, these methods have quite high accuracy, especially between similar languages. In [3], dynamic programming technique was used which allows the system to consider all possible alignments and thus find the minimum cost alignment. The method got a 4% error rate. A similar algorithm was developed in [9], where word counts were used instead of character counts. In [13], the same algorithm was applied to a corpus of English and Cantonese. The results were comparable.

The second approach, referred to as location-based approach, also depends on statistical information within the texts. In [8], sentences were aligned by using cognates (words that are similar phonetically across languages) at the level of character sequences. The algorithm developed in [6] aims at working on 'roughly parallel' texts (texts with certain sections missing in one language) and with unrelated language pairs. The method infers a small dictionary that helps the alignment.

The third approach used in sentence alignment is called the lexical approach, where the lexical information about texts are considered. Usually, a bilingual corpus is utilized to match the content words between the texts and these matches are used as anchor points. In [5], the algorithm starts by assuming that the first and last sentences of the texts align and they are the initial anchors. A variation of this idea was implemented in [12], with two basic differences. The function words were eliminated using a pos tagger and an online dictionary was used to find the matching word pairs. In another study, a simple word-to-word translation model was constructed and the best alignment was determined as the one that maximized the likelihood of generating the corpus [1].

3. COMPILING TURKISH-ENGLISH BILINGUAL CORPUS

An important goal of the research in this paper was forming a reliable and comprehensive bilingual corpus between Turkish and English languages. This was deemed as an important task, since such a resource enables researchers to develop corpus-based machine translation applications among these two languages. To this aim, we have carefully examined several types of resource of different genre, eliminated those that cannot be used for the intended purpose, and formed a thorough classification of the texts with respect to some criteria important for future applications. Below we list and give the details of the sources used for collecting the bilingual texts:

- *E-books:* These are electronic versions of some popular books (novels, stories, politics, etc.). Especially, we have made use of the Project Gutenberg, which made accessible old, popular and classical books in digital environment with the purpose of free access for readers [11]. In addition, the Turkish translations of some of the e-books were found in forum sites.
- *Articles in news sites:* Some Turkish newspapers also well-known abroad keep an English version of their websites. In these websites, the articles of some authors in the newspapers are periodically translated into English. These texts are very good sources since they are smaller and thus it is easier to trace the translation pattern used in the texts. These texts are translations from Turkish to English.

- *Academic work*: Most of these are formed of information texts, advertisements, and theses recorded in websites of academic units. We classified some of these into the group of technical data sources.
- *Documents from translation companies:* Translation companies can be regarded as keeping formal bilingual material. We have collected several documents from such companies. However, since these documents mostly have private content, all such information (company names, money amounts, etc.) in them should be cleaned before making them public. Since this is a time-consuming and error-prone process, we decided not to include these documents in corpus for the time being.

The contents of the corpus are shown in Table 1. Most of the column headings are self explanatory. The *category*, *type*, and *sub-type* fields are used to classify the entries according to their contents. The *quality* field is an indication of the translation quality between the Turkish and English versions. The field was assigned one of three values (very good, good, adequate) after a careful examination. Although it is not easy to determine the quality exactly, such a classification is necessary since it is a usual practice for alignment algorithms to measure their performances on texts with different qualities. There are a few additional fields in the classification table, omitted here.

4. THE SENTENCE ALIGNMENT ALGORITHM AND THE RESULTS

The algorithm developed in this research is a combination of location-based and length-based sentence alignment approaches. Given the two texts, first the texts are divided into paragraphs and sentences. Though paragraph identification can be done with a very low error rate, sentence identification poses more difficulties. There are several algorithms for sentence splitting [2,4,7,10]. We have used the LingPipe splitter. The method we propose is formed of two phases working in a similar manner. In the first phase, the paragraphs in the source and target texts are aligned. In the second phase, for each paragraph pair, the sentences within the paragraphs are aligned. Both types of alignment follow the same logic. In the case of paragraph alignment, initially all the paragraphs in the texts are considered and for each possible source and target paragraph pairs, a score is calculated. Then the pair with the minimum score is aligned, provided that the score is less than a threshold value. Following this, both texts are divided into two parts: the paragraphs above the aligned ones and those below the aligned ones. Then the algorithm is called recursively for these two sub-documents.

The score corresponding to the pair "ith source paragraph and jth target paragraph" is calculated using the following equation:

$$(1) \qquad score_{i,j} = \alpha_{i,j}\left(\frac{up-s_i}{up-t_j}-\beta\right)^2 + \left(\frac{len-s_i}{len-t_j}-\beta\right)^2 + \left(\frac{dn-s_i}{dn-t_j}-\beta\right)^2$$

where

$$\alpha_{i,j} = \frac{\dfrac{len-s}{len-s_i}+\dfrac{len-t}{len-t_j}}{2} \qquad \beta = \frac{len-s}{len-t}$$

Len-s and len-t denote the source text length and target text length, respectively; len-s_i and len-t_j denote the length of the ith source paragraph and the length of the jth target paragraph, respectively; up-s_i and up-t_j denote the length of the source text above the ith paragraph and the length of the target text above the jth paragraph, respectively; dn-

Tab. 1: Contents of Turkish-English bilingual corpus.

Id	Name	Category	Type	Sub Type	Word No	Page No	Quality
B001	Harry Potter – Philosopher's Stone	Book	Novel	Fantasy	56,000	170	Good
B002	Harry Potter – Chamber of Secrets	Book	Novel	Fantasy	67,000	189	Good
B003	Harry Potter – The Prisoner of Azkaban	Book	Novel	Fantasy	84,600	178	Good
B004	Harry Potter – Goblet of Fire	Book	Novel	Fantasy	150,000	302	Good
B005	Harry Potter - The Order of the Phoenix	Book	Novel	Fantasy	200,000	418	Good
B006	J.R.R. Tolkien - The Lord of the Rings : The Fellowship of the Ring	Book	Novel	Fantasy	142,000	450	Good
B007	J.R.R. Tolkien - The Lord of the Rings : The Two Towers	Book	Novel	Fantasy	119,000	380	Good
B008	J.R.R. Tolkien - The Lord of the Rings : The Return of the King	Book	Novel	Fantasy	106,000	310	Good
B009	George Orwell – 1984	Book	Novel	Science fiction	65,000	220	Good
B010	W. Shakespeare - Macbeth	Book	Play	Drama	18,200	32	Adequate
B011	Stephen King - Pet Sematary	Book	Novel	Horror	87,000	142	Good
B012	Dan Brown - The Da Vinci Code	Book	Novel	Police story	77,200	295	Good
B013	Descartes - Discourse on Method	Book	Philosophy	Politics	24,700	47	Good
B014	Bacon - New Atlantis	Book	Philosophy	Politics	13,000	33	Good
B015	Plato - Statesman	Book	Philosophy	Politics	18,700	108	Good
B016	Tommaso Campanells - City of Sun	Book	Philosophy	Politics	23,700	40	Good
B017	Dostoyevski – Notes from the Underground	Book	Novel	Drama	30,000	98	Good
B018	Henry D.Thoreau – Resistance to Civil Governement	Book	Philosophy	Politics	8,300	21	Good
B019	Tolstoy - Anna Karenina	Book	Novel	Drama	351,000	883	Good
B020	Aristoteles – The Athenian Constitution	Book	Philosophy	Politics	24,400	43	Good
B021	Plato – Republic	Book	Philosophy	Politics	43,400	349	Good
B022	Mark Twain - Tom Sawyer	Book	Novel	Adventure	71,000	139	Very good
B023	Voltaire – Candide	Book	Novel	Drama	36,600	80	Good
B024	Carl Von Clausewitz - War	Book	Study	War	98,000	105	Adequate
B025	Lenin – The State and Revolution	Book	Study	Politics	28,900	90	Good
B026	Plato – Apology	Book	Study	Politics	11,600	42	Good
B027	Cicero – Treatises on Friendship and Old Age	Book	Philosophy	Study	22,000	65	Good
B028	Stephen King - Green Mile	Book	Novel	Romance	134,000	443	Good
B029	Carus - On the Nature of Things	Book	Philosophy	Drama	74,000	175	Adequate
B030	Tolstoy - Master and Man	Book	Novel	Drama	19,200	64	Good
B031	Tolstoy - Ivan Ilic	Book	Novel	Drama	15,800	32	Good
B032	David Eddings – The Belgariad : Pawn of Prophecy	Book	Novel	Fantasy	79,540	157	Adequate
B033	David Eddings – The Belgariad : Queen of Sorcery	Book	Novel	Fantasy	106,000	195	Adequate
B034	David Eddings – The Belgariad : Magician's Gambit	Book	Novel	Fantasy	97,000	180	Adequate
B035	David Eddings – The Belgariad : Castle of Wizardry	Book	Novel	Fantasy	120,000	206	Adequate
B036	David Eddings – The Belgariad : Enchanter's End Game	Book	Novel	Fantasy	116,580	197	Adequate
B037	Arthur Clarke – 2001 A Space Odyssey	Book	Novel	Science fiction	61,850	138	Good
B038	Arthur Clarke – Rama II	Book	Novel	Science fiction	114,470	245	Good
B039	Arthur Clarke - Rendezvous with Rama	Book	Novel	Science fiction	72,000	193	Good
B040	Bernard Shaw - Caesar and Cleopatra	Book	Play	Drama	39,000	102	Good
B041	Kafka - Metamorphosis	Book	Story	Drama	15,700	28	Adequate
B042	Goethe - Faust	Book	Poetry	Drama	12,700	40	Adequate
B043	Gogol - Taras Bulba	Book	Novel	Drama	51,760	94	Good
B044	Eleanor H.Porter - Pollyanna	Book	Novel	Drama	95,000	301	Very good
B045	Anatole France - Thais	Book	Novel	Adventure	36,600	69	Good
B046	Dostoevsky – The Brothers Karamazov	Book	Novel	Drama	350,000	562	Good
B047	Ivan Turgenev - Rudin	Book	Novel	Drama	53,460	118	Good
B048	Robert L.Stevenson - Markheim	Book	Story	Drama	5,600	11	Good
B049	Dostoyevski – The Gambler	Book	Novel	Drama	62,850	126	Good
B050	Goethe - Iphigenia in Tauris	Book	Play	Drama	19,630	45	Very good
B051	Lermontov - A Hero of Our Time	Book	Novel	Drama	37,000	68	Good
B052	Moliere - The Imaginary Invalid	Book	Play	Critique	14,900	61	Adequate
B053	G. Leroux -Mystery of Yellow Room	Book	Novel	Police story	47,250	85	Good
B054	Jack London - The Call of the Wild	Book	Novel	Adventure	33,600	63	Good
B055	Dostoyevski - Devils	Book	Novel	Politics	260,000	440	Adequate
B056	Balzac - Eugenie Grandet	Book	Novel	Drama	55,750	93	Good
B057	Balzac - Hidden Masterpiece	Book	Story	Drama	13,300	27	Good
B058	Anatole France - Penguin Island	Book	Novel	Adventure	52,800	91	Very good
B059	Chamisso - Peter Schlemihl	Book	Novel	Psychology	38,360	75	Good
B060	Oscar Wilde-The Happy Prince and Other Tales	Book	Story	Kid	10,700	18	Very good
B061	Dostoevsky - Crime and Punishment	Book	Novel	Psychology	203,000	330	Good
T001	Bilkent University – Core Regulations	Short text	Regulations		2,800	7	Adequate
T002	Erhan Sigorta – Jewellers Block Insurance	Short text	Policy		3,300	9	Very good
T003	Boğaziçi University - New Approach in Courses	Short text	Regulations		2,440	7	Adequate
T004	Boğaziçi University - Graduate Record	Short text	Mail		345	1	Good
T005	Plesk Server	Short text	Technology		499	2	Very good
T006	Working Capital Handbook	Short text	Guide		3,200	10	Very good
T007	News	Short text	Article		61,880	125	Adequate
T008	Hotels Manual	Short text	Advertisement		432	2	Adequate
T009	The Turkish National Anthem	Short text	Poetry		101	1	Adequate
T010	Ninni	Short text	Story		485	2	Good
T011	Şeftali	Short text	Story		358	2	Good
T012	Inscribed Rock	Short text	Advertisement		147	1	Good
T013	Martial Dances	Short text	Poetry		207	1	Adequate
T014	Friend, You're not the Guilty One	Short text	Poetry		73	1	Adequate
T015	Children Love One Another	Short text	Poetry		44	1	Adequate
T016	Do not Forget	Short text	Poetry		57	1	Adequate
T017	The Triangle of Existence	Short text	Poetry		88	1	Adequate
XOO1	Subtitles of Movies	Subtitle			64,545	457	Adequate
X002	University Theses	Thesis			2,426	13	Very good

s_i and dn-t_j denote the length of the source text below the ith paragraph and the length of the target text below the jth paragraph, respectively. Note that β represents the ratio of the lengths between the source and target texts, and the overall score tends to be small when the source and target paragraphs reside in positions with nearly equal distances from the beginning and end of the corresponding documents.

After the score is calculated for each pair of paragraphs between the two texts, the pair with the minimum score is selected. If this score is less than a threshold value, then the paragraphs are aligned and the algorithm continues with alignment of the upper and lower parts of the paragraphs just aligned. In case that the minimum score exceeds the threshold value, it is considered that the paragraphs cannot be aligned in a 1-1 fashion and the whole set of paragraphs in the source and target range are aligned. As the paragraph alignment is completed, the sentence alignment phase begins. For each pair of source and target paragraphs aligned, the sentences within them are aligned independent of the other parts of the documents. The same formula is used, with the modification of replacing paragraph lengths with sentence lengths.

Two points about the method are worth noting. First, we do not use a predefined threshold value, instead the threshold value changes dynamically according to the size of the text portions to be aligned. The larger the size of this portion, the higher the value of the threshold. For instance, if the source and target parts contain only a few number of sentences, then the threshold value is small and we require an alignment as accurate as possible. The second point is that the proposed method allows alignment schemes other than 1-1, such as 1-2, 2-1, 2-3 alignments. This is a quite common situation in translated texts, especially in the case of sentence alignment.

The proposed method was applied on some of the documents listed in Section 3, in order to test the success of the method and the quality of the corpus. Due to lack of space, we here give the results on only three of these documents. The documents with different characteristics were selected in order to observe the performance of the algorithm on different types of document. The results are displayed in Table 2.

Document 1 is a text containing long paragraphs in both languages and having somewhat similar paragraph counts. But it is a hard text when we consider the sentence alignment beads. The percentage of 1-1 beads is only 65.2% and the percentage of 1-2 or 2-1 beads is 22.3%. The remaining pairs consist of more complex beads, even 1-6, 1-5 or 2-5. It also contains a deleted region of 18 sentences long in English text which is hard to handle. Under these situations, the algorithm did 63% of the alignments correctly and 24% were complete errors. The remaining 13% was partial errors: for instance, the real bead is 1-2, but the algorithm splits it into two beads as 1-1 and 0-1. Another important point is the question of how much the deleted block affected the overall performance. The 18-sentence long segment was towards the end of the text. For a short period during execution, it caused the algorithm to give continuous wrong alignments. But it managed to overcome this situation later. When we exclude this segment, the accuracy increases to 73.7%, which is quite high for such a difficult text.

In the second experiment, we obtained low success rates. The paragraph alignment phase outputed several wrong matches, since there were a large number of 1-6, 1-5, etc. paragraph beads. When the algorithm failed in paragraph alignment, it inevitably made errors in sentence alignment in large blocks. Due to this problem, the accuracy was about 45%. The last experiment was performed on a document where most of the paragraph correspondences were 1-1. Also, in the sentence level, 1-1 bead percentage was high (about 90%). Under these values, the algorithm resulted a very good

Tab. 2: Performance of the algorithm.

	No of sentences	1-1 rate	No of paragraphs	1-1 rate	Success %	Partial error %	Complete error %
Document 1	330 / 440	65.2	49 / 51	93.0	63.0	13.0	24.0
Document 2	153 / 146	86.0	69 / 22	25.0	45.0	10.0	45.0
Document 3	138 / 142	89.0	38 / 37	90.0	96.1	2.2	1.7

accuracy. The percentage of correct alignments was 96.1% and 2.2% was partial alignment errors. Only 1.7% of all alignments was completely wrong.

5. CONCLUSIONS AND FUTURE WORK

In this research, we formed a comprehensive bilingual Turkish-English corpus, developed a sentence alignment method, and tested the proposed method on the compiled corpus. To the best of our knowledge, this is the most comprehensive bilingual corpus among these two languages. The corpus and its wide coverage can serve as an important data resource for machine translation applications by the researchers.

The developed algorithm, like most sentence alignment algorithms, performs better for texts with well-arranged paragraphs. A future enhancement can be on increasing the robustness of the algorithm so that it can give comparable results on other types of text. Another issue is about missing segments in one of the documents. Since the algorithm works location-based, it takes some time to recover after a missing segment. In future work, we plan to shorten the length of this recovery period by using lexical information.

6. REFERENCES

[1] Chen, S.F. (1993) *Aligning sentences in bilingual corpora using lexical information*. Proc. of 30th Annual Meeting of ACL, 9-16.

[2] Finch, S., Mikheev, A. (1997) *A workbench for finding structure in texts*. Proc. of 5th Conference on Applied Natural Language Processing, 372-379.

[3] Gale, W.A., Church, K.W. (1991) *A program for aligning sentences in bilingual corpora*. Proc. of the 29th Annual Meeting of ACL, 177-184.

[4] GATE Framework. http://gate.ac.uk/.

[5] Kay, M., Röscheisen, M. (1994) Text-translation alignment. *Computational Linguistics* 19:1, 121-142.

[6] Li, W., Liu, T., Wang, Z., Li, S. (1994) *Aligning bilingual corpora using sentences location information*. Proc. of 3rd ACL SIGHAN Workshop, 141-147.

[7] LingPipe. http://www.alias-i.com/lingpipe/.

[8] Melamed, I.D. (1996) *A geometric approach to mapping bitext correspondence*. IRCS Technical Report 96-22, University of Pennsylvania.

[9] Moore, R.C. (2002) Fast and accurate sentence alignment of bilingual corpora. *Machine translation: from research to real users*. Springer-Verlag, 135–244.

[10] Nadeau, D. (2005) *Balie: baseline information extraction*. Technical Report, School of Information Technology and Engineering, University of Ottawa.

[11] Project Gutenberg. http://www.gutenberg.org/.

[12] Simard, M., Plamondon, P. (1998) Bilingual sentence alignment: balancing robustness and accuracy. *Machine Translation* 13:1, 59–80.

[13] Wu, D. (1994) Aligning a parallel English-Chinese corpus statistically with lexical criteria. Proc. of the 32nd Annual Meeting of ACL, New Mexico, 80–87.