Application of a Word-Alignment Algorithm to Bilingual Greek-Latin Documents

DIONYSIOS SOTIROPOULOS [1,2] , ELENI GALIOTOU[1] , CHRISTOS SKOURLAS[1]

[1]Department of Informatics
TEI of Athens
Ag. Spyridona, 12210 Egaleo
GREECE

[2]Department of Informatics
University of Piraeus
18534 Piraeus
GREECE

http://www.cs.teiath.gr

Abstract: - In this paper we address the problem of improving the access to bilingual Greek – Latin documents by incorporating their word level parallelization into the search process. More specifically we focus our experimentation on finding word correspondences between the Greek and Latin versions of the Gospel of Matthew by the utilization of the K – vec algorithm. We demonstrate the applicability of the algorithm in this broader class of language pairs by measuring its efficiency in identifying word correspondences that are valid mutual translations. Our implementation and evaluation of the K – vec algorithm for this specific language pair verifies that the 75% of the top ranked candidate word correspondences produced by the algorithm are valid mutual translations.

Key-Words: - parallel corpora, word alignment, statistical language processing, bilingual document processing, greek-latin text

1 Introduction

The motivation for the application of a word alignment algorithm lies in an attempt to improve the access and search the content of bilingual historical and religious documents.

The fact that the search process is conducted on a word level basis imposes the need for aligning the bilingual texts at the word level. Our primary concern is to provide a software tool that allows the user to search for a specific keyword in one of the languages pertaining to the specific language pair. However, the search results returned to the user should cover all instances of the selected keyword in the given language and all the instances of the corresponding keyword in the other language.

Thus, the choice of an appropriate alignment algorithm at the word level lies within the core of our system since it is of crucial importance to identify valid mutual translations.

This paper is organized as follows: section 2 describes the framework of our experimentation, while section 3 elaborates on the choice of the K – vec algorithm. Section 4 briefly discusses the underpinnings of the selected algorithm and focuses on demonstrating its effectiveness in identifying valid word correspondences between Greek and Latin bilingual religious documents. Specifically, the experimental results presented in that section quantify the applicability of K – vec to this specific language pair.

Finally, we summarize the paper and draw conclusions in section 5.

2 The Framework

The experimentation with a statistical alignment algorithm is motivated by the need to improve the access to the content of historical religious bilingual documents following the methodology used in the "Politimo" project. The project aims at the development and use of innovative techniques of image processing, pattern recognition and natural language processing for the development of an open architecture information system for the processing, management and accessing to the content of valuable collections of historical books and manuscripts.

Our experimentation is performed on a fragment of the "Politimo" corpus, which consists of bilingual Greek-Latin religious documents such as the Gospel of Matthew.

These documents share the following characteristics:

- They are not aligned at a sentence level
- No reliable lexical cues are available

- No orthographic / phonetic cognates are available

These characteristics were taken into account for the choice of the appropriate algorithm as described in the following section.

3 Word Alignment – The choice of the K – vec algorithm

The existing approaches to the problem of aligning bilingual documents at the word level can be divided into two categories.

The first category involves those algorithms that rely on prealigned texts at the sentence level. Wu and Xia [2] introduced a hybrid statistical and lexical alignment strategy in order to learn a bilingual lexicon from the aligned data that could be embedded within the learned translation models. Their approach utilized a dynamic programming algorithm that provided an optimized approximation to the word level alignment of English and Chinese texts. The bilingual training process was based on an iterative EM (expectation maximization) procedure for maximizing the likelihood of generating the Chinese corpus given the English text. The output of the training process was a set of potential Chinese translations for each English word, together with the probability estimate for each translation.

However, this methodology depends largely on very small set of high – reliability lexical cues, which had to be known in advance and were not available for the parallel documents of this project.

Matsumoto, Ishimoto and Utsuro [3] described a method for finding structural matching between parallel sentences of two languages. Parallel sentences in their approach were analyzed based on unification grammars, and making use of a similarity measure of word pairs in the two languages performed structural matching.

Although their approach serves as a useful source for extracting linguistic and lexical knowledge it could not be applied for the word level parallelization of the documents of this project. This is mainly due to the fact that the historical documents pertaining to the corpus of “Politimo” are not aligned at the sentence level. Secondly, the required similarity measure between word pairs of Greek and Latin assumes some a priori lexical knowledge that was not available.

Marcu and Wong [4] presented a joint probability model for machine translation, which automatically learns word and phrase equivalents form bilingual corpora. The translation model adapted in their approach assumed that word correspondences can be established not only at the word level, but at the phrase level as well. Additionally, their model does not try to capture how source sentences can be mapped into target sentences, but rather how both sentences can generated simultaneously. Nevertheless, this algorithm could not be applied for the purposes of "Politimo" since the available parallel documents in the corpus were not aligned at the sentence level.

All the previous approaches rely on the existence of prealigned texts at the sentence level. Traditional algorithms performing such a task, Brown, Lai and Mercer [5], Gale and Church [6], exploit the correlation between the lengths of mutual translations on the character or the token level. These approaches formulate the problem of obtaining a valid alignment between the given bilingual texts as a ML (maximum likelihood) estimation problem. Specifically, the only feature influencing the probability of their alignment is a function of their differences in their lengths in characters or tokens. However, the empirical estimation of the distributions of these probabilities requires the availability of an adequate number of hand – aligned training bitexts.

The second category consists of those algorithms whose operation does not depend on the existence of sentence aligned parallel texts. In this context the Smooth Injective Map Recognizer (SIMR) proposed by Melamed [7] formulates the bitext-mapping problem in terms of pattern recognition. SIMR infers bitext maps form likely points of correspondence between the two texts, points that are plotted in a two – dimensional space of probabilities.

The most important difficulty in adapting the SIMR approach for the purposes of "Politimo" arises from the fact that the functionality of the algorithm is based on the existence of orthographic / phonetic cognates or translation lexicons. Two words, coming from languages that share the same alphabet, are considered orthographic cognates when they have the same meaning and similar spellings.

When dealing with language pairs that have different alphabets phonetic cognates represent words that have similar sounds. However, such information was not available in the context of “Politimo”.

The choice of the K – vec algorithm introduced by Fung and Church [1] was made on the grounds of its simplicity since it relies on the fact that if two words are translations of each other they occur almost an equal number of times and approximately in the same region of the parallel text. Moreover, it does not require any prior knowledge of the

punctuation or the sentence boundaries of the languages under consideration. Thus, it is an open question to explore the applicability of the algorithm by demonstrating its efficiency in a broader class of language combinations. A similar approach was followed by Utsuro et al. [8] where they applied statistical techniques in order to estimate word correspondences.

4 Alignment of the Bilingual Greek - Latin Text

The K – vec algorithm identifies word pairs that might be mutual translations by noticing the distribution of each word in the corresponding text. For example, the word pair (mercedem, μισθόν) constitutes a valid mutual translation pair which the algorithm manages to identify by taking into consideration only the distribution of "mercedem" in the Latin text and the corresponding distribution of "μισθόν" in the Greek text.

The K – vec algorithm estimates these word occurrence distributions by partitioning both texts into an equal number of text fragments and measuring the frequency of each word in the corresponding fragments. Subsequently, by utilizing statistical measures such as the Dice's Coefficient, the algorithm estimates which word pairs are more likely to be mutual translations.

Table 1

Pair ID	Latin Word	Estimated Greek Word	Corrected Greek Word
1	totum	ολον	
2	mercedem	μισθον	
3	manum	ηψατο	χείρα
4	adimpleretur	πληρωθη	
5	iohannes	Ιωάννης	
6	obtulerunt	προσήνεγκαν	
7	bethleem	Βηθλέεμ	
8	omnem	πασαν	
9	synagogis	συναγωγαις	
10	adimpleretur	ρηθεν	
11	puerum	παιδίον	
12	et	και	
13	in	και	
14	caelorum	ουρανων	
15	autem	δε	
16	potest	δύναται	
17	omnia	πάντα	
18	ioseph	Ιωσηφ	
19	adimpleretur	λέγοντος	πληρωθή
20	audistis	Ηκούσατε	
21	matrem	μητέρα	
22	caelorum	βασιλεία	ουρανών
23	meum	μου	
24	angelus	Ιωσηφ	άγγελος
25	istis	τούτων	
26	et	ο	και
27	dico	λέγω	
28	ecce	ιδου	
29	in	δε	εν
30	est	και	εστί
31	ego	εγω	

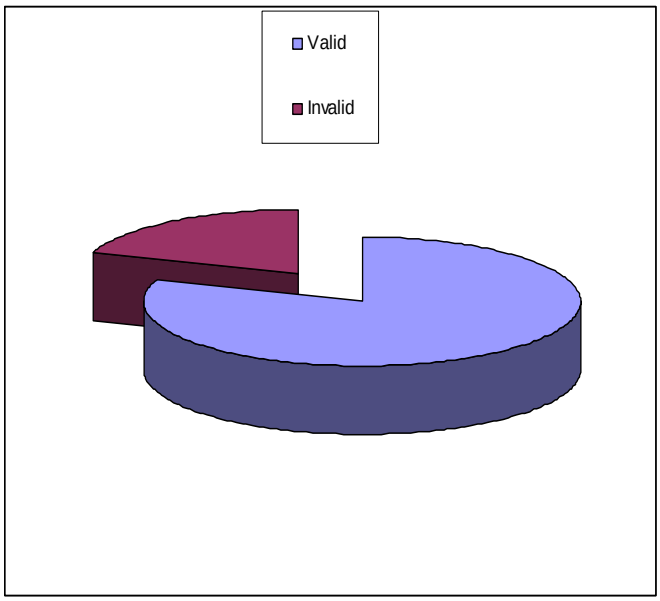

Figure 1

Tables 1 and 2 summarize the evaluation results obtained from the execution of the K – vec algorithm on the Greek and Latin versions of the Gospel of Matthew. Our evaluation verifies that up to the 75 percent of the top ranked estimated mutual translations (24 out of 31) as is is shown in Figure 1 are indeed valid word correspondences.

Particularly, Table 1 presents a list of the top ranked word correspondences estimated by K – vec. The second column of the table contains the Latin words of each pair while the third column contains the estimated Greek translations. Finally, the fourth column contains the corresponding corrections on the Greek translations in those cases where the algorithm failed to identify a valid mutual translation pair.

On the other hand Table 2 exhibits all the statistical related information for each pair of the estimated mutual translations. Specifically, the second column lists the Dice's Coefficient values for each word pair presented in Table 1. The values N1, N2 and N3 listed in the third, fourth and fifth

columns contain word occurrence frequency information. Specifically, N1 represents the total number of text fragments containing both words of each pair. N2 represents the total number if text fragments containing only the Latin word while N3 constitutes the total number of pairs containing only the Greek word.

Thus, the Dice's coefficient for each word pair appearing in Tables 1 and 2 may be computed as shown in (1):

$$Dc = \frac{2 * N1}{N2 + N3} \quad (1)$$

Table 2

Pair ID	Dice Coefficient	N1	N2	N3
1	1.000	4	4	4
2	1.000	5	5	5
3	1.000	4	4	4
4	1.000	5	5	5
5	1.000	5	5	5
6	1.000	4	4	4
7	1.000	4	4	4
8	1.000	4	4	4
9	1.000	5	5	5
10	1.000	5	5	5
11	1.000	4	4	4
12	0.9859	70	72	70
13	0.9481	64	65	70
14	0.9412	8	9	8
15	0.9107	51	53	59
16	0.9091	5	6	5
17	0.9091	5	6	5
18	0.9091	5	6	5
19	0.9091	5	5	6
20	0.8889	4	5	4
21	0.8889	4	5	4
22	0.8889	8	9	9
23	0.8889	4	5	4
24	0.8889	4	4	5
25	0.8889	4	4	5
26	0.8661	55	72	55
27	0.8649	16	18	19
28	0.8571	15	19	16
29	0.8548	53	65	59
30	0.8430	51	51	70
31	0.8421	8	11	8

5 Conclusions

In this paper, we have presented the results of the application of a word-alignment algorithm on bilingual Greek-Latin documents. In particular, we have applied the K-vec algorithm [1] on a Greek-Latin version of the Gospel of Matthew. We have demonstrated the applicability of the algorithm by measuring its efficiency in a series of experiments with a language pair such as Greek and Latin that do not share a common alphabet.

References:

[1] P. Fung, K. Church, "K – vec A new approach for Aligning Parallel texts", *Proc. 15th International Conference on Computational linguistics,* Kyoto, Japan, 1994, pp. 1096 – 1102.

[2] D. Wu, X. Xia, "Learning an English Chinese lexicon from parallel corpus, *In Amta-94, Association for Machine Translation in the Americas,* Columbia, 1994, pp. 206-213

[3] Y. Matsumoto, H. Ishimoto and T. Utsuro, "Structural Matching of Parallel Texts", *In Proc. 31st Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, June 1993, pp. 23 – 30.

[4] D. Marcu, W. Wong, "A Phrase – Based, Joint Probability Model for Statistical Machine Translation", *Proc. of the Conference on Empirical Methods in Natural Language Processing,* Philadelphia, PA, 2002, pp. 133–139.

[5] P. F. Brown, J. C. Lai and R. L. Mercer, "Aligning Sentences in Parallel Corpora", *In 29th Annual Meeting of the Association for Computational Linguistics,* Berkeley, California, 1991, pp. 169 – 176.

[6] W. A. Gale, K. W. Church, "A program for aligning sentences in bilingual corpora", *Computational Linguistics*, vol. 19, no. 1, pp.75-102.

[7] I. D. Melamed, "Bitext maps and alignment via pattern recognition", *Computational Linguistics,* vol. 25, no. 1, pp.107–139.

[8] T. Utsuro, H. Ikeda, M. Yamane, Y. Matsumoto and M. Nagao, "Bilingual text matching using bilingual dictionary and statistics.", *In Proc. Of COLING'94,* pp. 1076–1082.

Acknowledgements

This work was co-funded from the E.U. by 75% and from the Hellenic State by 25%, through the Operational Program «Information Society» - Measure 3.3. «Research and Technological Development in the Information Society».