ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Automatic Alignment of Medical vs. General Terminologies

Laura Dioşan[1,2], Alexandrina Rogozan[1] and Jean Pierre Pécuchet[1]

[1]LITIS (EA 4051), INSA, Rouen, France
[2]Babeş-Bolyai University, Cluj-Napoca, Romania

Abstract. We propose an original automatic alignment of definitions taken from different dictionaries that could be associated to the same concept although they may have different labels. The alignment between a specialized terminology used by the librarians to index concepts and a general vocabulary employed by a neophyte user in order to retrieve documents on Internet, will certainly improve the performances of the information retrieval process. The selected framework is a medical one. We propose a terminology alignment by an SVM classifier trained on a compact, but relevant representation of such definition pair by several similarity measures and the length of definitions. Three syntactic levels are investigated: Nouns, Nouns-Adjectives, and Nouns-Adjectives-Verbs. Our aim is to show how the combination of similarity measures offers a better semantic access to the document content than only one measure and it improves the performances of the automatic alignment. The results obtained on the test set show the relevance of our approach, as the F-measure reaches 88%. However, this conclusion should be validated on larger corpora.

1 Introduction

One of the most important characteristic of an information retrieval system is related to its capability to answer queries of both neophyte and expert users. The expert user queries are formulated in a specialized language, which is generally the language used to index the documents for a very particular domain, as the health area. The problem is that the neophyte users formulate their queries with a naive language, while the documents are indexed through the concepts of specialised terminologies. Therefore, it becomes necessary to automatically align several specialised terminologies with the vocabulary shared by an average user for information retrieval on Internet. These alignments will allow the information retrieval systems for a better exploitation of specialised terminologies and electronic dictionaries in order to benefit from the advantages of their strengths.

Our aim is to achieve an accurate automatic alignment of medical definitions in French taken from several specialised terminologies with those from general dictionaries. This alignment is a difficult task since these definitions may have different labels, although they are related to the same medical concept. Therefore, we have chosen to represent the specialised terminology by several concepts taken from the thesaurus Medical Subject Headings (MeSH) and the VIDAL dictionary, while the medical general vocabulary is represented by several definitions from the encyclopaedia Wikipedia and from the semantic network of *Le*

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Dictionaire Integral (LDI), which is an ensemble of linguistic resources provided by Memodata[1]. Therefore, the concept alignment is actually viewed in terms of definition alignment. The main aim is to design an algorithm that given two definitions (expressed as text sentence(s)) will decide whether they refer to the same concept or not. In order to perform this alignment, each definition, corresponding to a given concept and taken from a dictionary, is first turned into a bag of words with some semantic labelling; then a pair of bags of words corresponding to two definitions is turned into a standard feature vector in a low dimensional space (R^7) via some similarity measures; the obtained vectors are used as observations for an SVM classifier that will decide if the two definitions are aligned or not (in other words if they refer to the same concept).

Aligning two definitions actually means to solve a binary classification problem. A representative corpus of aligned and non-aligned definitions has been created in order to allow the classifier to learn the discrimination of such relationships and then to evaluate the performance of our approach for the automatic alignment of definitions[2].

We decide to perform a linguistic analysis of definitions by using several natural language processing techniques as segmentation, lemmatisation, and syntactic labelling in order to obtain richer and more robust descriptors than simple strings of characters, making thus a more relevant definition matching possible. Moreover, the definitions are considered at three syntactic levels: the level of nouns (N), the level of nouns and adjectives (NA) and the level of nouns, adjectives, and verbs. These levels allow us to measure the contribution of each syntactic form to the performance of the alignment system. *A priori*, we do not know which are the most important: the nouns, the adjectives or/and the verbs.

This paper is structured as follows: first, the related work for the sentence alignment is presented. The linguistic processing we perform for each definition is summarised in Section 3. Our model for definition alignment is detailed in the next section. Before concluding, several numerical experiments are presented and discussed.

2 Related work

To our knowledge, only the problem of aligning sentences from parallel bilingual corpora has been intensively studied for automated translation. While much of the research has focused on the unsupervised models [1, 2, 3], a number of supervised discriminatory approaches have been recently proposed for automatic alignment [4, 5, 6]. One of the first algorithms used to align parallel corpora proposed by Brown [1] and developed by Gale [2] is based solely on the number of words/characters in each sentence. Chen [3] has developed a simple statistical word-to-word translation model. Dynamic programming is used to perform the

[1] `http://ist.inserm.fr/basismesh/mesh.html`, `http://www.vidal.fr/`, `http://www.memodata.com/2004/fr/dicologique/index.shtml`, `http://www.memodata.com`

[2] The set of definitions has been achieved by G. Lortal, I. Bou Salem and M. Wang during the VODEL project

search for the best alignment in these models.

Concerning the supervised methods, Taskar et al. [4] has cast the word alignment as a maximum weighted matching problem in which each pair of words in a sentence pair is associated with a score, which reflects the desirability of the alignment of that pair. Moore [5] has introduced a hybrid and supervised approach that adapts and combines the sentence-length-based methods with the word-correspondence-based methods. Ceausu [6] has proposed another supervised hybrid method that uses an SVM classifier to distinguish between aligned and non-aligned examples of sentence pairs, each pair has being represented by a set of statistical characteristics. Related to the use of linguistic information more recent work [7] shows the benefit of combining multilevel linguistic representations. Moreover, data fusion has been exhaustively investigated in the literature, especially in the framework of IR [7, 8].

The definition alignment is a different and more difficult problem. The parallelism of corpora refers actually to the meaning of the content, which is expressed in the same language, but using different vocabularies. Moreover, the problem associated with a classical representation TFIDF from bags of words, which involves very large sparse input vectors must be avoided. Therefore, we propose a new representation that allows a fast and efficient learning of definition alignment.

3 Our corpora and the linguistic processing

The aim of the current research is to align the definitions from the electronic dictionaries in order to establish the correspondences between a specialised terminology and a general one. A representative corpus has been created from an user vocabulary and a medical one in order to allow the intermediation of two terminologies. The medical thesaurus MeSH is used in order to index the health documents, as those belonging to the CISMeF health catalogue, whereas the VIDAL dictionary is destined especially to the patients and their families. The LDI dictionary used by Memodata represents the knowledge shared by non-experts. The LDI does not cover a specific domain, but it contains a large set of concepts used by a neophyte user in a natural language. To complete this area, we considered a set of medical concepts with their definitions from the encyclopaedia Wikipedia. The French is the common language for all the definitions. This provides us with six data sets, which represent pairs of sentences taken from two different dictionaries.

Several linguistic treatments have to be considered in order to improve the performances of the automatic alignment. The segmentation consists in cutting a sequence of characters such as to bring together various characters that form a single word. We choose to cut the sequences of characters depending on several separation characters such as “space”, “tab” or “backspace”. The lemmatisation is the process for reducing inflected and even derived words to their stem, base or root form. The stem has not to be identical to the morphological root of the word; it is usually sufficient that related words map the same stem, even if this

stem is not in itself a valid root. The syntactic labelling affixes the corresponding syntactic label, such as noun, adjective or verb to each word. This allows us to filter the empty words (such as the punctuation signs) and to consider only those that are pregnant with meaning.

Moreover, the syntactic labelling led us to the comparison of the automatic definition alignment performance at three different syntactic levels: one that retains only the nouns from each definition (the N level), one that retains only the nouns and the adjectives from each definition (the NA level) and another one that retains the nouns, the adjectives and the verbs from each definition (the NAV level). This allows us to obtain a bag of words representation that is precise and meaningful.

4 The proposed model

Recently, an unsupervised automatic alignment model (denoted by *uS1*) have been implemented based only on a similarity measure chosen among several well-known ones (as *Matching*, *Dice*, *Jaccard*, *Overlap* or *Cosine* coefficient) [9]. Moreover, a supervised alignment by an SVM algorithm (denoted by *sS1*), which takes into account only a similarity measure, has been also investigated. The performances of these automatic alignments have shown that it is no possible to identify only one measure that provides the best alignments for all the dictionary combinations (see Table1).

Therefore, a new framework for the automatic alignment of definitions, which takes into account the complementarities between these similarity measures, is investigated in the current paper. This time all five similarity measures are simultaneously considered. In addition to these measures, the length of each definition is taken into account. The model (denoted by *sS5*) is a hybrid one because it combines an algorithm that computes the similarity measures between two definitions with an SVM classifier [10], which decides, based on these measurements, if two definitions are aligned or not. The SVM algorithm uses an RBF kernel in order to discriminate such relationship. The output is represented by the class (aligned or not aligned) to be associated to each couple of definitions.

The parameters of our model (the penalty for miss-classification C and the bandwidth σ of the RBF kernel) are optimized by a parallel grid search method in a cross-validation framework in order to avoid the over fitting problems. Thus, we automatically adapt the SVM classifier to the problem, actually the alignment of definitions.

5 Experiments

Several numerical experiments are performed for six different combinations of dictionaries and for different syntactic levels: N, NA and NAV. Our supervised model (*sS5*) is compared with an unsupervised model (*uS1*) [9] and with an SVM-based supervised model (*sS1*), the last two models taking into account

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

only a similarity measure. Note that for each syntactic level, only the best similarity measure among several well-known measures is retained– see Table 1.

The C-SVM algorithm, provided by LIBSVM [11], with an RBF kernel is used in the experiments. The optimisation of the hyper parameters is performed by a parallel grid search method in the following ranges: the tuning coefficient C is optimised in the $[2^{-10}, 2^{10}]$ range; the bandwidth σ of the RBF kernel is optimised in the interval $[2^{-4}, 2^{1}]$; a 10-fold cross validation is performed during the training phase. Several performance measures, borrowed from the information retrieval domain, are used in order to evaluate the automatic alignment we propose: the precision of alignments, the recall of alignments, and the *F-measure* - the weighted harmonic mean of precision and recall. The confidence intervals are computed as well.

	Memo vs. MeSH	*Memo vs. Vidal*	*Memo vs. Wiki*	*MeSH vs. Vidal*	*MeSH vs. Wiki*	*Vidal vs. Wiki*	
uS1	40.0±13.34	48.0±13.58	50.0±13.59	46.0±13.55	52.0±13.58	63.0±13.12	
sS1	79.55±3.24	79.55±3.24	81.40±3.13	78.65±3.29	78.16±3.32	79.07±3.27	N
sS5	80.95±3.16	81.54±3.12	84.27±2.93	81.39±3.13	79.6±3.24	87.35±2.67	
uS1	44.0±13.49	54.0±13.55	50.0±13.59	48.0±13.58	54.0±13.55	56.0±13.49	
sS1	72.73±3.58	34.15±3.81	74.42±3.51	32.79±3.77	75.29±3.47	77.11±3.38	NA
sS5	77.10±3.38	73.58±3.54	85.71±2.81	80.00±3.21	84.78±2.89	87.99±2.61	
uS1	44.0±13.49	52.0±13.58	52.0±13.58	46.0±13.55	42.0±13.42	48.0±13.58	
sS1	31.88±3.74	33.80±3.80	20.69±3.25	79.07±3.27	67.42±3.77	71.26±3.64	NAV
sS5	77.10±3.38	79.06±3.27	85.39±2.84	80.00±3.21	80.00±3.21	84.08±2.94	

Table 1: F measures and their confidence intervals of different alignment models.

The results on the test set show the relevance of our supervised SVM-based approach, because the F-measures of *sS5* are larger than those of *uS1* and *sS1* in all cases. The model *uS1* considers the fact that for our data sets one definition has to be aligned with another definition. We do not have this bias for the supervised models (*sS1* and *sS5*) because by using these classifier-based models it is possible to align a definition from a dictionary with any definitions, with one definition or with several definitions from the other dictionary, respectively.

The model based on learning from nominal groups like NA (Nouns-Adjectives) seems to lead to the best performances. Concerning the NAV model, better results are obtained for the combination Memodata *vs.* Wikipedia only (both dictionaries using a general language). In the other cases, the verbs do not improve the performance of the alignment system. However, these conclusions should be checked on larger corpus.

6 Conclusion and further work

This paper has presented a new model for the automatic alignment of definitions taken from general and specialised dictionaries. The definition alignment

has been considered as a binary classification problem and an SVM algorithm has solved it. In order to achieve this aim the classifier has used a representation of definitions based on several similarity measures and the definition lengths that is compact and pertinent. The definitions have been considered at three syntactic levels and the influence of each level has been analysed. The information conveyed by the nouns and adjectives seem to be more relevant than those from the verbs are. However, these conclusions should be validated for large corpora.

Further work will be also focused on developing a definition alignment based on a bag of morpho syntactic patrons. A representation of definitions enriched by semantic and lexical extensions (synonyms, hyponyms, and antonyms) will also be considered.

Acknowledgment

The VODEL RNTL-ANR Grant has supported this research.

References

[1] P. F. Brown et al. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1994.

[2] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. In *ACL*, pages 177–184, 1991.

[3] S. F. Chen. Aligning sentences in bilingual corpora using lexical information. In *MACL*, pages 9–16, 1993.

[4] B. Taskar, S. Lacoste-Julien, and D. Klein. A discriminative matching approach to word alignment. In *HLT '05*, pages 73–80. ACL, 2005.

[5] R. C. Moore. Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02*, pages 135–144. Springer-Verlag, 2002.

[6] A. Ceausu, D. Stefanescu, and D. Tufis. Acquis communautaire sentence alignment using SVM. In *LREC*, pages 2134–2137, 2006.

[7] F. Moreau, V. Claveau, and P. Sébillot. Automatic morphological query expansion using analogy-based machine learning. In G.Amati et al., editor, *ECIR 2007*, volume 4425 of *LNCS*, pages 222–233. Springer, 2007.

[8] W. B. Croft. *Combining approaches to information retrieval*, chapter 1, pages 1–36. Kluwer Academic Publishers, 2000.

[9] G. Lortal et al. Du terme au mot : Utilisation de techniques de classification pour l'alignement de terminologies. In *TIA*, 2007. accepted.

[10] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[11] C-C. Chang and C-J. Lin. *LIBSVM: a library for support vector machines*, 2001. (`http://www.csie.ntu.edu.tw/~cjlin/libsvm`).