Mutual terminology extraction using a statistical framework

Le An Ha
University of Wolverhampton
L.A.Ha@wlv.ac.uk

Ruslan Mitkov
University of Wolverhampton
R.Mitkov@wlv.ac.uk

Gloria Corpas
Universidad de Malaga
gcorpas@ya.com

Abstract: In this paper, we explore a statistical framework for mutual bilingual terminology extraction. We propose three probabilistic models to assess the proposition that automatic alignment can play an active role in bilingual terminology extraction and translate it into mutual bilingual terminology extraction. The results indicate that such models are valid and can show that mutual bilingual terminology extraction is indeed a viable approach.

Resumen: El presente trabajo aborda la utilización de un marco estadístico para la extracción de terminología bilingüe por asociación o información mutua. Se proponen tres modelos probabilísticos para evaluar si el alineamiento automático puede desempeñar un papel activo en la extracción de terminología bilingüe y si ello es extrapolable a la extracción de terminología bilingüe por información mutua. Los resultados indican que dichos modelos son válidos y que la extracción de terminología bilingüe por información mutua puede ser un enfoque viable.

Keywords: Automatic terminology extraction, bilingual terminology extraction.

1 Introduction

The identification of terms in scientific and technical documents is a crucial issue for any application dealing with analysis, understanding, generation, and translation of such documents. Throughout the last decade, computational linguists, translators, lexicographers, and computer engineers among other specialists have been interested in automatically identifying terminology in texts. Software tools to accomplish terminologically related tasks have been designed and implemented. There is also increasing interest in bilingual terminology extraction (BLTE) (detailed in Section 2), whose usual approach is monolingual terminology extraction followed by automatic alignment. Recently, it has been suggested that automatic alignment can play a bigger role in bilingual terminology extraction, by assuming that if a noun phrase in the target language is aligned to a term in the source language, this noun phrase is more likely to be a term (Ha et al. 2008). In that paper, the authors provide an ad hoc framework to assess the effect of the term scores of the source language noun phrases on the term scores of the target language noun phrases. Whereas such an ad hoc assessment is a promising approach, it does rely on experiments to find the optimised settings.

In this paper, we will provide a statistic framework to examine this assumption by providing several models in which the probability of a noun phrase to be a term in a target language is affected by the probability of its alignment to a term in the source language. Our statistical models, therefore, provide a better foundation for mutual bilingual terminology extraction.

This paper is organised as follows: After the introduction (this section), we will discuss terminology extraction in general (Section 2). Our models are then presented in Section 3. Evaluation of the proposed models can be found in Section 4. Conclusions and future directions are found in Section 5.

2 Terminology extraction (monolingual and bilingual)

2.1 Monolingual terminology extraction

The main stages in terminology work can be summarised as: extraction of terms from a corpus, validation of terms found, and organisation of validated terms by domain and sub-domain (Sauron, 2002). In this respect, a number of projects have created automatic extraction tools, which identify term candidates by starting from a corpus in electronic form. Some projects go one step further: on the basis of parallel corpora of texts and their translations they propose not only candidate terms but also possible equivalents in a target language.

Approaches to term extraction (TE) are usually classified as linguistic, statistical, or hybrid. Linguistic and statistical approaches can be further subdivided into term-based (intrinsic) and context-based (extrinsic) methods (cf. Bourigault et al., 2001; Streiter et al., 2003).

Terminology Extraction tools (TETs) following a linguistic approach try to identify terms by their linguistic (morphological and syntactic) structure. For this purpose, texts are annotated with linguistic information with the help of morphological analysers, part-of-speech taggers and parsers. Then, term candidates (TCs) following certain syntactic structures are filtered from the annotated text by using pattern matching techniques. Intrinsic methods try to filter TCs according to their internal (i.e morphological) structures (Ananiadou 1994). Extrinsic methods, on the other hand, try to identify TCs by analysing the morpho-syntactic structure of a word or phrase, such as looking for part-of-speech sequences like NP= noun + noun (e.g. computer science). An example of this kind is represented by the program LEXTER (Bourigault, 1992). Another commonly used technique consists in filtering TCs by looking for commonly used text structures such as definitions and explanatory contexts like “X is defined as …” or “X is composed of …” (cf. Pearson, 1998).

The general assumption underlying the statistical approaches to TE is that specialised documents are characterised by the repeated use of certain lexical units or morpho-syntactic constructions. TETs based on statistics try to filter out words and phrases having a certain frequency-based statistic higher than a given threshold (see Manning & Schütze 1999 for an overview). Another common method is to compare the frequency of words and phrases in a specialised text to their frequency in general language texts, assuming that terms tend to appear more often in specialised texts than in general language texts.

Different evaluation criteria exist for TETs, involving among others accuracy, as well as supported file formats and languages. The most frequently used criteria are noise and silence, as well as recall and precision. While noise refers to the ratio between discarded TCs and the accepted ones, silence refers to the number of terms not detected by a TET. Recall and precision are two measures frequently used in IR, the former being defined as the ratio between the number of correctly retrieved terms and the number of existing terms, the latter being defined as the ratio between correctly extracted terms and the number of proposed TCs (cf. Zielinski, 2002).

TETs following a purely linguistic approach tend to produce too many irrelevant TCs (noise), whereas those following a purely statistical approach tend to miss TCs that appear with a low frequency value (silence, cf. Clematide, 2003). Linguistic-based TETs often provide better delimited TCs than statistical-based ones. However, the disadvantage of linguistically based TETs is that they are language-dependent and thus only available for major languages. Statistical TETs, on the other hand, can be used for lesser-used languages that lack computational resources such as minority languages (cf. Streiter et al., 2003).

More recently, approaches to automatic TE and TR have moved towards using both statistical and linguistic information (Daille et al., 1994; Justeson & Katz, 1996; Frantzi, 1998). Generally, the main part of the algorithm is the statistical part, but shallow linguistic information is incorporated in the form of a syntactic filter which only permits phrases having certain syntactic structures to be considered as candidate terms.

2.2 Bilingual terminology extraction

Most of what has been discussed so far applies to monolingual TE and TR. Lately, research has evolved towards the automatic extraction of bilingual terms. This process involves automatically capturing bilingual terminology from existing technical texts and their translations (parallel corpora), validating the candidate term pairs generated and generating terminological records in an automatic or semi-automatic manner. Several works have focused on the extraction of knowledge from bilingual corpora. All of them address the problem of aligning units across languages. Although very successful methods have been designed to align paragraphs and sentences in two different languages, aligning units smaller than a sentence still raises a real challenge.

Thus, Gaussier (1998) relies on corpora aligned at the sentence level. Association probabilities between single words are calculated on the basis of bilingual co-occurrences of words in aligned sentences. Then these probabilities are used to find the French equivalents of English terms through a flow network model. Hull (1998) differs from Gaussier (1998) in that single-word alignment, term extraction and term alignment are three independent modules. Terms and words are aligned through an algorithm that scores the candidate bilingual pairs according to probabilistic data, chooses the highest scored pair, removes it from the pool, and repeatedly recomputes the scores and removes pairs until all the pairs are chosen. Further improvements on Gaussier's first model can be found in Gaussier et al. (2000) and Dejean et al. (2003).

Chambers (2000) describes a project launched in 1999 whose main aims include the automatic capture of bilingual terminology from parallel corpora, the manual validation of bilingual term pairs and the automatic generation of terminological records. The whole process has three major operations: monolingual extraction in the source text, monolingual extraction in the target text and bilingual matching to produce candidate term pairs.

Many methods have been proposed for extracting translation pairs from bilingual corpora, but most are based on word frequency and are, therefore, not effective in extracting low-frequency pairs. Word-frequency-based methods are language-pair-independent. Examples include Melamed (2000) and Hiemstra (1997). While popular and well-known translation pairs may already be included in existing bilingual dictionaries, newly coined and minor translation pairs are not very well-covered in available resources. In order to tackle this problem, Tsuji & Kageura (2004) present a method for extracting low-frequency translation pairs from Japanese-English bilingual corpora. Their method uses transliteration patterns that are observed in actual loan-word pairs, thus incorporating language-pair-dependent knowledge.

More recently (Ha et al. 2008), it was proposed the use of automatic term alignment to help propagate the strengths of terminology extraction from one language into another. The availability of parallel corpora aligned at sentence level makes the alignment process more accurate, and thus makes this possible. The overall process of the mutual bilingual terminology extraction methodology can be described as follows: firstly, a list of term candidates is extracted for the first language; then term candidates from the second language are aligned to this list. If a term candidate in the second language is aligned to a term candidate in the first language, its term score is increased, and the candidate is promoted. This process can be repeated many times. In this study, as no suitable mathematical framework was employed, different settings had to be experimented with, in order to choose the best ones. To overcome this weakness, we propose in this paper several probabilistic models which can be used to propagate the term scores of a noun phrase in the source language to its aligned noun phrase in the target language.

3 Mutual bilingual terminology extraction

3.1 Three probabilistic models

Let $P(N_s)$ is the probability of the noun phrase N_s in the source language is a term, $P(N_t)$: the probability of the noun phrase Nt in the target language is a term, and $P(N_t=N_s)$ is the probability of the noun phrase N_s translated into the noun phrase N_t. Let $P_m(N_s)$ and $P_m(N_t)$ are the probabilities of the noun phrase Ns and Nt to be a term in monolingual contexts.

We will use the notion "model 0" to refer to automatic terminology extraction in the monolingual context. In the model 0, $P(N_s)=P_m(N_s)$ and $P(N_t)=P_m(N_t)$.

In model 1, we assume that the probability of the noun phrase Nt in the target language as a term only depends on whether it is a translation of a term in the source language, or in other words: $P(N_t)$=P(N_s is a term and N_s is translated into N_t). As N_s is a term and N_s is translated into Nt are two independent events, $P(N_t)$ is calculated as:

$$P(N_t)=P(N_s)*P(N_t=N_s)= P_m(N_s)*P(N_t=N_s). \quad (1)$$

This model is similar to the approach suggested by Gaussier (1998). This approach assumes that the target language only plays a passive role in terminology extraction.

In the next model (model 2), we assume that the probability of the noun phrase Nt in the target language to be a term does not only depend on whether it is a translation of a term in the source language, but also whether it is a term in the target language in the context of monolingual terminology processing. In this model, $P(N_t) = P($[N_t is a term in the target language or [N_s is a term in the source language and N_s is translated into N_t]). As [N_t is a term in the target language] and [N_s is a term in the source language and N_s is translated into N_t] are two overlapping, but independent events, the probability of the joint event is calculated as

$P(N_t)=P_m(N_t)+P(N_s)*P(N_t=N_s)-P_m(N_t)*P(N_s)*P(N_t=N_s)$ (2)

in which $P_m(N_t)$ is the probability of N_t is a term in the target language in a monolingual context.

In the third model (model 3), we propose that the probability of a noun phrase Ns in the source language as a term is also affected by the probability of its translation to be a term in the target language. In this way $P(N_s)$ in (2) should be calculated as

$P(N_s)=P_m(N_s)+P(N_t=N_s)*P(N_t)-P_m(N_s)*P(N_t=N_s)*P(N_t)$ (3)

(3) is a recursive formula: as $P(N_t)$ is calculated using $P(N_s)$ also. As a result, (3) should be rewritten as:

$P_0(N_s)=P_m(N_s)$

$P_0(N_t)=P_m(N_t)$

$P_{n+1}(N_s)=P_n(N_s)+P(N_t=N_s)*P_n(N_t)-P_n(N_s)*P(N_t=N_s)*P_n(N_t)$

$P_{n+1}(N_t)=P_n(N_t)+P(N_t=N_s)*P_n(N_s)-P_n(N_t)*P(N_t=N_s)*P_n(N_s)$

The calculation should be repeated until converged.

3.2 Calculating component probabilities

In the previous section, we proposed three different probabilistic models to calculate the probability of a noun phrase Nt to be a term given the probability of it being the translation of a noun phrase Ns, and the probability of Ns to be a term. The next step is to figure out how the probability of Ns to be a term can be calculated in the monolingual context. As discussed in Section 2, statistical measures have been derived to calculate the "termhood" of a term candidate. Although the value of these termhood functions is related to the probability of a noun phrase to be a term (i.e. the higher the value is, the more likely that it is a term), the actual probability is not often explicitly calculated. In order to calculate these probabilities using a known termhood function, we have to use linear regression as described below.

Given that $F(N)$ is a termhood function of the noun phrase N, $C(F(N))$ is the number of noun phrases N_i having $F(N_i)>=F(N)$, $T(F(N))$ is the number of confirmed terms T_i whose $F(T_i)>=F(N)$. The probability of a noun phrase N to be a term in a monolingual context can be estimated as:

$P_m(N)=T(F(N))/C(F(N))$

Our task is to find a function $G(F(N))$ which can be used as a good estimation of $T(F(N))/C(F(N))$. In order to find such a function, a graph between $F(N)$ and $T(F(N))/C(F(N))$ can be drawn. Figure 1 shows the relation between $F(N)$ and $T(F(N))/C(F(N))$ ($P_m(N)$) when $F(N)$ is calculated as the log of frequency of N, for 400 noun phrases in English found in a parallel corpus of English and Spanish Law (See Section 4). The graph indicates that $T(F(N))/C(F(N))$ (i.e. $P_m(N)$) and $log(Fre(N))$ seems to have a linear relation whose coefficients can be estimated using linear regression. (Assuming the relationship is $y=ax+b$, in this case, a=0.384 and b=-0.064).

The use of linear regression also has another benefit: the standard errors from the linear regression can also be used for estimating the predictive powers of termhood functions. A small standard error indicates that the termhood function is a good indicator of the probability of a noun phrase to be a term and vice versa.

We have experimented with several termhood functions, and it proves that frequency remains a very good termhood function (i.e it produces the smallest standard error when linear regression is used).

Having established how to calculate $P_m(N)$, we now move to calculate the probability of the noun phrase Ns translated into the noun phrase Nt. Using the sentence-aligned parallel corpus, we can use contingency tables to estimate this probability by employing log likelihood calculation (Manning and Schütze 1999).

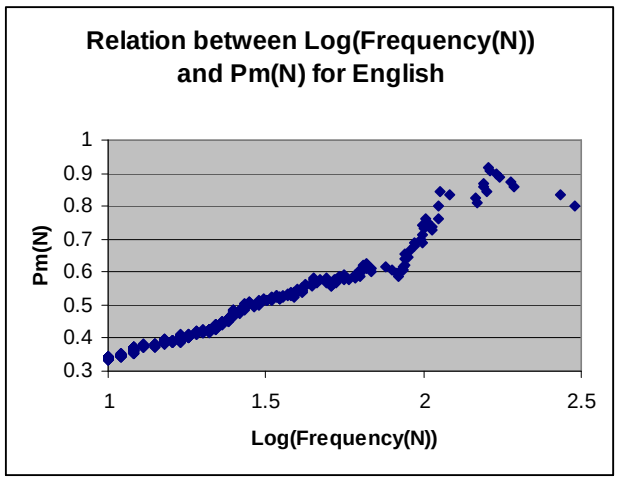

Figure 1: Relation between a termhood function and $P_m(N)$

4 The experiment

We compile a parallel corpus in the domain of EU Immigration law in English and Spanish. The corpus contains 4390 segments, 121534 English words and 136585 Spanish words. We use the Inter Active Terminology for Europe (IATE) as an authoritative source to confirm whether a noun phrase is a term in the domain or not. This confirmation is done for both English and Spanish.

In order to evaluate the three models, we calculate the standard errors of the predicted probability suggested by each model and the maximum likelihood probability of a noun phrase having the predicted probability greater than or equal the current one to be a term. This seems to be an unusual way to evaluate performance of automatic terminology extraction, but given that our main objective is to evaluate our probabilistic models, this is an appropriate choice: the smaller the standard error is, the better the model at predicting the probability of a noun phrase to be a term.

Table 1 shows the standard errors calculated as described above using the three proposed models (see Section 3), in which *Log(Frequency)* is used to estimate the initial probability of a noun phrase to be a term in the monolingual context. The results indicate that out of the three models, model 1 provides the most errors, whereas model 3 is slightly better than model 2. This confirms our mathematical prediction.

The results also indicate that weaknesses in term extraction in one language can be overcome by employing a corpus aligned at sentence level. In our case, the part-of-speech sequence pattern used for Spanish is not as good as the pattern used for English, resulting in a higher standard error in model 0 (monolingual terminology extraction) for Spanish. When mutual bilingual terminology extraction is applied, the standard errors have been reduced to much closer to that of English.

	Spanish	English
Model 0	0.056	0.026
Model 1	0.053	0.044
Model 2	0.04	0.024
Model 3	0.035	0.022

Table 1: Standard errors between predicted probability and maximum likelihood probability

In order to show our probabilistic models also work with other types of termhood function, we tried another type of combination in which $F(N_s)=Log(Fre(N_s))*Length(N_s)$ (in which $Length(N_s)$ is the number of words N_s has). $F(N_t)$ is still $Log(Fre(N_t))$. The results are shown in Table 2. These results also indicate that out of the three models, model 3 gives the most accurate prediction of the probability of a noun phrase to be a term. Nevertheless, it is shown that our models can also propagate weaknesses as well as strengths: the use of a less accurate termhood function in English can result in higher standard errors in Spanish. Other experiments in which different combination of $F(N_s)$ and $F(N_t)$ have been used have been performed. None of these experiments yield better results (in term of standard errors) when compared to the results given in Table 1.

	Spanish	English
Model 0	0.056	0.044
Model 1	0.065	0.054
Model 2	0.059	0.042
Model 3	0.057	0.04

Table 2: Standard errors when $F(N_s)=Log(Fre(N_s))*Length(N_s)$

5 Conclusions and future directions

In this paper, we propose three probabilistic models to incorporate alignment scores in automatic term extraction. The proposed probabilistic models have advantages over the Ha et al. (2008) approach in that they are built on sound mathematical basis, and the remaining problem shifts to calculating the probability of a noun phrase to be a term in the monolingual context, rather than performing different experiments to find an optimised way to normalise and incorporate different termhood functions. Using this approach, any termhood function can be used, if the function can be converted into a probabilistic function predicting the possibility of a noun phrase to be a term.

In the future, we will explore different ways to calculate the alignment probability, and propose new models to account for the fact that a term in the source language may have multiple translations.

Reference

Ananiadou, S. 1994. A methodology for Automatic Term Recognition. In Proceedings of the 15th International Conference on Computational Linguistics (COLING94), pp. 1034-1038. Kyoto, Japan.

Bourigault, D., C. Jacquemin, and M. C L'Homme (ed.) 2001. Recent Advances in Computational Terminology. Amsterdam: John Benjamins Publishing Company.

Chambers, D. 2000. Automatic Bilingual Terminology Extraction: A Practical Approach. In Proceedings of *Translating and the Computer 22, Aslib/IMI*.

Daille, B., E. Gaussier; J.-M. Lange. 1994. Towards

Automatic Extraction of Monolingual and Bilingual Terminology. In Proceedings of COLING 1994.

Dejean, H., E. Gaussier, C. Goutte, and K. Yamada. 2003. Reducing parameter space for word alignment. In Proceedings of *HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta.

Frantzi, K. T. 1998. Automatic Recognition of Multi-Word Terms. PhD Thesis. Manchester Metropolitan University, UK.

Gaussier, E. 1998. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In Proceedings of *Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pp. 444--450. San Francisco, California.

Gaussier, E., D. Hull, and S. At-Mokthar. 2000. Term alignment in use: Machine-aided human translation. In J. Veronis (ed.). Parallel text processing: Alignment and use of translation corpora, pp. 253--274. Dordrecht: Kluwer Academic Publishers.

Ha, L. A., G. Fernandez, R. Mitkov, and G. Corpas. 2008. Mutual bilingual terminology extraction. To appear in LREC 2008.

Hiemstra, D. 1997. Deriving a bilingual lexicon for cross language information retrieval. In Proceedings of Gronics 1997, pp. 21-26.

Hull, D. 1998. A practical approach to terminology alignment. In Proceedings of CompuTerm 1998, pp. 1-7.

Justeson, J. S., and S. L. Katz. 1996. Technical Terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 3(2): 259-289.

Manning, C. D., and H. Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Melamed, I. D. 2000. Models of translational equivalence among words. *Computational Linguistics* 26(2): 221-249.

Pearson, J. 1999. *Terms in context*. Amsterdam: John Benjamins.

Sauron, V. A. 2002. Tearing out the terms: evaluating terms extractors. In Proceedings of Translating and the Computer 24, London, Britain.

Streiter, O., D. Zielinski, I. Ties, and L. Voltmer. 2003. Term extraction for Ladin: An example-based approach. In Proceedings of TANL 2003 Workshop on Natural Language Processing of Minority Languages with few computational linguistic resources, Batz-sur la Mer.

Tsuji, K., and K. Kageura. 2004. Extracting low-frequency translation pairs from Japanese-English bilingual corpora. In Proceedings of CompuTerm 2004, pp. 23-30.

Zielinski, D., and Y. R. Safar. 2005. t-survey 2005: An Online Survey on Terminology Extraction an Terminology Management. In Proceedings of Translating and the Computer 27, London, Britain