Mining for Term Translations in Comparable Corpora

Dan Ştefănescu
Research Institute for Artificial Intelligence, Romanian Academy
Calea 13 Septembrie, No. 13, Bucharest 050711, ROMANIA
E-mail: danstef@racai.ro

Abstract

This paper presents the techniques currently developed at RACAI for extracting parallel terminology from the comparable collection of Romanian and English documents collected in the ACCURAT project. Apart from being used for enriching translation models, parallel terminology can be (and very often is) a goal in itself, since such resources can be used for building dictionaries or indexing technical or domain-restricted documents.

Keywords: Terminology Extraction, Terminology Mapping, Comparable Corpora

1. Introduction

The construction of any Statistical Machine Translation System requires two types of statistical models: language models and translation models, whose parameters are usually derived from the analysis of parallel corpora. However, large parallel corpora are only available for a quite small number of languages with rich resources (English, French, German, Spanish, etc.) and so, there is an increasing need in gathering parallel data for under resourced languages. One of the recent approaches in solving this task is to extract parallel data from comparable corpora. Such corpora consist in documents covering the same topic or subject, using more or less parallel expressions, entities or terminology. For instance, one can easily find Wikipidia[1] or news articles which are examples of strongly and respectively weakly comparable documents. The goal is to extract, if possible, the existing parallel data and use it to enrich poor translation models.

This paper presents the techniques currently developed at RACAI for extracting parallel terminology from the comparable collection of Romanian and English documents in the ACCURAT project (Skadiņa et al., 2012). Apart from being used for enriching translation models, parallel terminology can be (and very often it is) a goal in itself, since such resources can be used for building dictionaries or indexing technical or domain-restricted documents.

First, the terminology is monolingually extracted, taking into consideration both single and multi-word terms, while in the second step the extracted terms are mapped based on string similarity and existing dictionaries. The methods described are language independent as long as language specific resources are provided. The paper is structured as follows: the next section presents the monolingual terminology extraction, while section 3 describes the terminology mapping. Experiments and results are presented in section 4. The paper ends with conclusions and references sections.

[1] http://en.wikipedia.org/wiki/Romania vs. http://ro.wikipedia.org/wiki/Rom%C3%A2nia (27.03.2012)

2. Terminology Extraction

Terminology extraction is the subtask of Information Extraction which refers to extracting terms from a given corpus, relevant to the genre / domain of the corpus. This task dates back to the 70s and it was most studied in the 90s. This latter period saw an explosion of various approaches (Schütze, 1998) based on raw frequency and part of speech filters (Dolby et al., 1973; Justeson and Kats, 1995), low variance in relative position for multi-word terms (Smadja, 1993), hypothesis testing and mutual information (Church and Hanks, 1989), likelihood ratios on assumed distributions (Dunning, 1993), inverse document frequency on assumed distributions (Church, 1995), finite-state automaton parsing (Grefenstette, 1994), full parsing (Bourigault, 1993; Strzalkowski, 1995), semantic analysis (Pustejovsky et al., 1993), etc. Recent work includes that of Park et al. (2002), who focus on all possible parts-of-speech terminology taking into account out-of-vocabulary words, Wong et al. (2007), who use a probabilistically-derived measure – *Odds of Termhood*, for scoring and ranking term candidates for term extraction, or Velardi et al. (2008), who see the Web as a huge corpus of texts that can be processed to create and update specialized glossaries.

While the existence of various commercially available terminology extraction tools[2] might suggest that this is a sufficiently studied problem, in practice, users complain about the amount of manual work required to filter out much of the terms returned by such systems[3].

Our solution makes a clear distinction between single-word and multi-word terms, since their identification and extraction is usually performed by using different approaches.

[2] http://www.translationzone.com/en/translator-products/sdlmultitermextract/ (27.03.2012)
http://www.e-kern.com/en/kern/translations/terminology/terminology-extraction.html (27.03.2012)
http://www.wordfast.net/ (27.03.2012)

[3] http://www.proz.com/forum/software_applications/96347-terminology_extraction_software.html (27.03.2012)

2.1 Single-word terminology extraction

We approached the task of single-word terminology extraction by improving Damerau's method (Damerau, 1993) as it has been reported to yield very good results (Schütze, 1998; Paukkeri et al., 2008). Damerau's approach compares the relative frequency in the documents of interest (user corpus – C_U) to the relative frequency in a reference collection (reference corpus – C_R). The original formula for computing the score of a word *w* is:

$$score(w) = \frac{f(w, C_U)}{|C_U|} \div \frac{f(w, C_R)}{|C_R|} \quad (1)$$

where $f(w, C)$ is the frequency of *w* in corpus *C*, and $|C|$ is the total number of words in *C*. One can immediately notice that the score for a word is calculated according to the likelihood ratios of occurring in both corpora (that of the user and the reference). The main idea is to compare the maximum likelihood estimates (MLE) computed on the user corpus to the ones on the reference corpus. Consequently, the reference corpus should be a large, balanced and representative corpus for the language of interest. Essentially, the MLE on such a corpus is equivalent with a unigram language model:

$$P_{MLE}(w) = \frac{f(w, C_R)}{|C_R|} \quad (2)$$

In practice, such models are usually used in information retrieval to determine the topic of documents. Thus, Damerau's formula works by comparing two unigram language models.

It has been proven however, that due to data sparseness, the unigrams language models constructed only by the means of MLE behave poorly and that a proper smoothing should be performed (Chen and Goodman, 1998). To do this, we employ a variant of Good-Turing estimator smoothing (Kochanski, 2006) :

$$P_{GT}(w) = \frac{f(w, C_R) + 1}{|C_R| + |V_R|} \cdot \frac{E(f(w, C_R) + 1)}{E(f(w, C_R))} \quad (3)$$

where V_R is the vocabulary (the unique words in C_R) and $E(n)$ is the probability estimate of the word to occur exactly *n* times.

Let us consider a slightly modified example from (Kochanski, 2006): let us say we have a (reference) corpus with 40,000 English words which contains only one instance of the word "*unusual*": $f(w, C_R) = 1$. Let us also say that the corpus contains 10,000 different words that appear once and so, $E(1)$ = 10,000 / 40,000, and that we have 5,500 words that appear twice, giving $E(2)$ = 5,500 / 40,000. Again, let us consider that the total number of the unique words in the corpus is 15,000 ($|V_R|$ = 15,000). The Good-Turing estimate of the probability of "*unusual*" is:

$$P_{GT}(unusual) = \frac{1+1}{40{,}000 + 15{,}000} \cdot \frac{5{,}500/40{,}000}{10{,}000/40{,}000}$$
$$= \frac{2}{55{,}000} \cdot \frac{5{,}500}{10{,}000} = \frac{1}{50{,}000}$$

But using MLE, we would have had a larger value:

$$P_{MLE}(unusual) = \frac{1}{40{,}000}$$

Because the sum of the probabilities must be 1, we have a remaining probability mass (P_R) to be reassigned to the unseen words (*U*). Consequently, for computing the estimated probability of a single unseen word *u_w*, we should divide this mass to the estimated number of unseen words $|U|$:

$$P_{GT}(u\_w) = \frac{P_R}{|U|} = \frac{E(1)}{(|C_R| + |V_R|) \cdot |U|} \quad (4)$$

Going back to Damerau's formula, we have now that:

$$score(w) = \frac{f(w, C_U)}{|C_U|} \div P_{GT}(w \text{ in } C_R) \quad (5)$$

The words having the highest scores are terminological terms. In case C_U is a large corpus, we can also compute Good Turing estimators for the numerator. For small corpora, this is however unreliable since one cannot compute the estimates $E(n)$ with high enough confidence.

This approach can be improved by additional preprocessing of the corpora involved. First, for better capturing the real word distribution, it is better to use word lemmas (or stems) instead of the occurrence forms. Second, the vast majority of the single terminological terms are nouns and therefore one can apply a Part of Speech (POS) filtering in order to disregard the other grammatical categories. Both can be resolved by employing stand-alone applications that can POS-tag and lemmatize the considered texts. As our research and development is mainly focused on English and Romanian, we usually make use of the TTL preprocessing Web Service (Ion, 2007; Tufiş et al., 2008) when dealing with these languages.

The method presented above can be reinforced with the well-known *TF-IDF* (*term frequency – inverse document frequency*) approach (Spärck Jones, 1972), provided that the corpus of interest is partitioned into many documents or that this partitioning can be automatically performed.

As reference corpora we used the *Agenda* corpus (Tufiş and Irimia, 2006) and a collection of *Wikipedia* documents for Romanian, while for English, we also used *Wikipedia* documents.

2.2 Multiple-word terminology extraction

Terminology extraction does not limit to single-word terms and so, one must be able to extract multi-word terminology, too. Smadja (1993) was among the first to advocate that low variance in relative position is a strong indicator for multi-word terminological expressions, which can be found among the collocations of a corpus. These are expressions which sometimes cannot be translated word-by-word using only a simple dictionary and a language model, because they might be characterized by limited compositionality – the meaning of the expression is more than the sum of the meaning of the words composing the collocation.

Different methods have been proposed for finding

collocations. Some counted the occurrences of bigrams and then used a part-of-speech filter in order to rule out those bigrams which cannot be phrases (Justeson and Krats, 1995). Smadja (1993) employed a method based on the mean and the variance of the distances between pairs of words, while others (Church et al., 1991) used *t Test*, *chi square Test*, *Log-Likelihood* or *Mutual Information* for finding pairs of words which appear together in the text more often than expected by chance.

Our approach for the identification and extraction of collocations has been described in several papers (Ştefănescu et al., 2006; Todiraşcu et al., 2009; Ştefănescu, 2010). For the purposes of the current task, we define a collocation as a pair of words for which:
– the distance between them is relatively constant;
– they appear together more often than expected by chance: *Log-Likelihood*.
Looking at this definition, one can notice, that from a strict linguistic point of view, such a construction can be seen as a strong co-occurrence, rather than a collocation.

The first component of our solution is based on a method developed by Smadja (1993). This uses the average and the standard deviation computed on distances between words to identify pairs of words that regularly appear together at the same distance, a fact which is considered to be the manifestation of a certain relation between those words. Collocations can be found by looking for such pairs for which standard deviation is small.

In order to find terminological expressions, we employ a POS filtering, computing the standard deviation for **only** the *noun-noun* and *noun-adjective* pairs within a window of 11 non-functional words length, and we keep all the pairs for which standard deviation is smaller than 1.5 – a reasonable value according to (Manning and Schütze, 1999). This method allows us to find good candidates for multi-word expressions but not good enough. We want to further filter out some of the pairs so that we keep only those composed by words which appear together more often than expected by chance. We do this by computing the Log-Likelihood (LL) scores for all the above obtained pairs, by taking into account only the occurrences of the words having the selected POS-es. We take into consideration the pairs for which the LL values are higher than 9, as for this threshold the probability of error is less than 0.004 according to the *chi square* tables.

We further keep as terminological expressions only those for which at least one of the words composing them can be found among the *single-word* terminological terms, disregarding their context. In this way we aim at filtering out commonly used expressions which have no terminological value.

3. Terminology mapping

Lately, automatic terminology mapping has been well-studied using methods like compositional analysis (Grefenstette, 1999; Daille and Morin, 2008) or contextual analysis (Fung and McKeown, 1997). Still, terminology mapping for languages with scarce resources is less researched (Weller et al., 2011).

Our terminology mapping tool was developed under the name TEA (TErminology Aligner). Given two lists containing monolingually extracted terminology, it is designed to find (in those lists) pairs of expressions which are reciprocal translations. In order to do this, TEA analyzes candidate pairs, assigning them translation scores (*tScore*) based on (i) translation equivalence estimation and (ii) cognates that can be found in those pairs (eq. 6).

$$tScore(pair) = max(te(pair), cg(pair)) \quad (6)$$

The translation equivalence score (*te*) for two expressions is computed based on the word-level translation equivalents existing in the expressions (eq. 7). Each word w_s in the source terminological expression e_s is paired with its corresponding word w_t in e_t such that the translation probability is maximal, according to a GIZA++ (Och and Ney, 2000) like translation dictionary.

$$te(e_s, e_t) = \frac{\sum_{w_s \in e_s} \max_{w_t \in e_t} dicScore(w_s, w_t)}{length(e_s) + \delta} \quad (7)$$

where *dicScore* is the translation equivalence score from the dictionary. The score should be normalized with the length of expression e_s. Still, we modify the denominator in order to penalize (δ) candidate pairs according to the length difference between source and target expressions:

$$\delta = \frac{|length(e_s) - length(e_t)|}{2} \quad (8)$$

The cognate score for two expressions is computed as the Arithmetic mean between two different string similarity measures (eq. 9). The first one (*sm_ld*) is calculated as the *Levenshtein Distance* (LD) in which the expressions are normalized (*norm*) by removing double letters and replacing some character sequences: "*ph*" by "*f*", "*y*" by "*i*", "*hn*" by "*n*" and "*ha*" by "*a*". This type of normalization is often employed by spelling and alteration systems (Ştefănescu et al., 2011). In practice, we modify this function in order to obtain values in the [0,1] interval, which we want to be high in case strings are similar and approach 0 for high differences (eq. 10). The second string similarity measure is simply the *longest common substring* of the two expressions, normalized by the maximum value of their lengths (eq. 11).

$$cg(e_s, e_t) = \frac{sm\_ld + sm\_lcs}{2} \quad (9)$$

$$sm\_ld = 1 - \frac{LD\big(norm(e_s), norm(e_t)\big)}{\min\big(length(e_s), length(e_t)\big)} \quad (10)$$

$$sm\_lcs = \frac{length(LCS(e_s, e_t))}{\max(length(e_s), length(e_t))} \quad (11)$$

The values of *te*(*pair*) and *cg*(*pair*) are taken into account only if they are higher than a threshold, the value of which regulates the tradeoff between precision and recall.

4. Experiments and Results

Evaluation of parallel terminology extraction requires the existence of a *Gold Standard* (GS) containing bilingual mapped terminology relevant to a collection of bilingual comparable texts. The only freely available such GS we know of is Eurovoc (Steinberger et al., 2002). This is "*the thesaurus covering the activities of the EU and the European Parliament in particular*" and it has been described in (Steinberger et al., 2002). We conducted two experiments: the first one was designed to assess the performance of the monolingual terminology extraction, while the second one, the performance of the mapping.

In the first experiment we considered 950 English-Romanian parallel documents from the *JRC-Acquis* corpus (Steinberger et al., 2006). They are all from 2006 and contain about 3.5 million tokens per language (approx. 55 Mb of preprocessed text). To assess the performance of the tool, we generated lists containing only those Eurovoc terms that appeared in these documents for both languages and counted how many of the recognized terms were found in these corresponding restricted lists (Table 1).

	English	**Romanian**
#documents	950	
Size of preprocessed	3.55 mil. tokens 55.1 Mb	3.34 mil. tokens 61.8 MB
Eurovoc terms identified out of those found in the collection having at least 1 occurrence	793 / 2699 29.38%	744 / 1961 37.93%
… 10 occurrences	289 / 1185 24.38%	252 / 815 30.92%
… 50 occurrences	65 / 507 12.82%	63 / 326 19.32
… 100 occurrences	24 / 318 7.54%	33 / 213 15.49%

Table 1: Eurovoc terms identified as terminological

If a word becomes more and more frequent, approaching its occurrence probability in the reference corpus, the tool cannot consider it terminological. This means, that some of the terminology that is valid for the entire JRC-Acquis cannot be discovered by considering only the documents from a single year, even though that terminology appears in those documents.

Regarding this evaluation methodology, one has to keep in mind that the list of Eurovoc terms is neither exhaustive nor definitive and as such, there might be valid non-Eurovoc terms that our application discovers. Examples for English include "*Basel convention*", "*standards on aviation*", "*Strasbourg*", "*national safety standards*", "*avian influenza*" etc. This is the reason for which we are not evaluating this module in terms of standard precision and recall.

For the second experiment, we considered the ideal case in which the monolingual terminology contains only and all the Eurovoc terms. We conducted this experiment for two language pairs: English-Romanian and English Latvian, computing precision (P), recall (R) and F-measure (F1) values. The next tables summarize the results.

Threshold	**P**	**R**	**F1**
0.1	0.563	0.069	0.122
0.2	0.426	0.101	0.163
0.3	0.562	0.194	0.288
0.4	0.759	0.295	0.425
0.5	**0.904**	**0.357**	**0.511**
0.6	0.964	0.298	0.456
0.7	0.986	0.216	0.359
0.8	0.996	0.151	0.263
0.9	0.995	0.084	0.154

Table 2: Terminology Mapping Performance for English-Romanian

Threshold	**P**	**R**	**F1**
0.1	0.347	0.068	0.114
0.2	0.357	0.108	0.166
0.3	0.636	0.210	0.316
0.4	0.833	0.285	0.425
0.5	**0.947**	**0.306**	**0.463**
0.6	0.981	0.235	0.379
0.7	0.996	0.160	0.275
0.8	0.996	0.099	0.181
0.9	0.997	0.057	0.107

Table 3: Terminology Mapping Performance for English-Latvian

We should mention that these ideal experiment settings, in which we deal with parallel data, allow us to assess the performance of our approach in situations which **can be compared** for the languages of interest. The described methodology for terminology identification is **monolingual** and therefore, it does not matter if the initial data is parallel, or merely comparable. The idea here is to allow for comparable scenarios. As the mapping process does not depend on the document collection, but only on the lists of monolingually extracted terms, again, it does not depend directly upon the comparability level of the initial data. In the mapping experiment described above, we were interested in the limit case where the extracted terminology can be entirely mapped. In the case of comparable corpora, the comparability level and the collection genres have both an important impact on the comparability of the monolingually extracted term lists. Accordingly, many terms may not be present in both lists and so, they cannot and should not be mapped. We might even end up with completely unmappable lists. This issue is the subject of further research.

5. Conclusions

This paper presents the techniques currently used for extracting parallel terminology from the comparable collection of Romanian and English documents in the ACCURAT project. The purpose of the this task is to improve the automatic alignment process of comparable corpora, which finally aims at developing better translation models for Statistical Machine Translation systems.

Future work will be focused on improving this approach by introducing a filtering step for eliminating some of the terms which are incorrectly found as terminological, as a consequence of the error propagation caused by the chaining of the statistical modules involved. We are also working on improving the evaluation process and on estimating the performance of our method for several other language pairs.

The mapping module is the basic terminology mapping tool in the ACCURAT project and it is currently involved in mapping terminology extracted for all the languages involved: English, Estonian, German, Greek, Croatian, Latvian, Lithuanian, Romanian and Slovenian.

6. Acknowledgements

This work has been supported by the ACCURAT project (http://www.accurat-project.eu/) funded by the European Community's Seventh Framework Program (FP7/2007-2013) under the Grant Agreement no. 248347.

7. References

Bourigault D. (1993). *An endogenous corpus-based method for structural noun-phrase disambiguation*, in Proceedings of EACL-93, pp. 81--86.

Chen S.F., Goodman J. (1998). *An empirical study of smoothing techniques for language modeling*, Technical Report TR-10-98, Harvard University.

Church K. (1995). *One term or two?*, in Proceedings of SIGIR-95, pp. 310--318.

Church K., Gale W., Hanks P., Hindle D. (1991).*Parsing, word associations and typical predicate-argument relations*, Current Issues in Parsing Technology. Kluwer Academic, Dordrecht.

Church K., Hanks P. (1989). *Word Association Norms, Mutual Information, and Lexicography*, in Proceedings of the 27th Annual Meeting of the ACL.

Daille, B. and Morin, E. (2008). *Effective Compositional Model for Lexical Alignment*. In Proceedings of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India.

Damerau F. (1993). *Generating and evaluating domain-oriented multi-word terms from text*. Information Processing and Management, 29(4), pp. 433--447.

Dolby J. L., Ross I.C., Tukey J. W. (1973, 1973, 1975, 1973). Index to Statistics and Probability, Vol. 1 The Statistics Cumindex, 2 Citation Index, 3-4 Permuted Title, 5 Locations and Authors. R and D Press.

Dunning T. (1993). *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics 19(1), pp. 61–74.

Fung, P. and McKeown, K. (1997). Finding terminology translations from non-parallel corpora.In Proceedings of the 5th Annual Workshop on Very Large Corpora, pp. 192--202.

Grefenstette G. (1994). *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Press, Boston.

Grefenstette, G. (1999). *The World Wide Web as a resource for example-based machine translation tasks*. Translating and the Computer 21, London, UK.

Ion R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD thesis (Romanian), Romanian Academy, Bucharest.

Justeson J.S., Katz S.M. (1995). *Technical Terminology: some linguistic properties and an algorithm for identification in text*. Natural Language Engineering (1), pp. 9--27. Cambridge University Press.

Kochanski G. (2006). *Lecture 4 - Good-Turing probability estimation*, Oxford.

Manning C., Schütze H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge.

Morin, E. and Prochasson, E. (2011). *Bilingual Lexicon Extraction from Comparable Cor-pora Enhanced with Parallel Corpora*. ACL HLT 2011, page 27.

Och, F. J. and Ney, H. (2000). *Improved Statistical Alignment Models*. In Proceedings of the ACL 2000, Hong Kong, China, pp. 440--447.

Park, Y., Byrd, R. J., Boguraev B. (2002). *Automatic glossary extraction: beyond terminology identification*. Proceedings of the 19th International Conference on Computational Linguistics - Taipei, Taiwan.

Paukkeri M., Nieminen I.T., Pöllä M., Honkela T. (2008). A *Language-Independent Approach to Keyphrase Extraction and Evaluation*, in Proceedings of COLING- 08.

Pustejovsky J., Bergler S., Anick P. (1993). *Lexical semantic techniques for corpus analysis*, Computational Linguistics, 19(2), pp. 331--358.

Schütze, H. (1998). *The Hypertext Concordance: A Better Back-of-the-Book Index*. In Proceedings of Computerm '98 (Montreal, Canada, 1998), D. Bourigault, C. Jacquemin, and M.-C. L'Homme, Eds., pp. 101--104.

Skadiņa, I., Aker, A., Glaros, N., Su, F., Tufiş, D., Verlic, M., Vasiļjevs, A., Babych, B. (2012). *Collecting and Using Comparable Corpora for Statistical Machine Translation*, in Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.

Smadja F. (1993). *Retrieving Collocations from Text: Xtract*.Computational Linguistics 19, pp. 143--175.

Spärck Jones, K. (1972). *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation 28 (1), pp. 11--21.

Ştefănescu D. (2010). *Intelligent Information Mining from Multilingual Corpora*, PhD thesis (Romanian), Romanian Academy, Bucharest.

Ştefănescu, D., Ion R., Boroş, T. (2011). *TiradeAI: An*

Ensemble of Spellcheckers, in Proceedings of the Spelling Alteration for Web Search Workshop, pp. 20--23, Bellevue, USA.

Ştefănescu, D., Tufiş, D., Irimia, E. (2006). *Automatic Identification and Extraction of Collocations from Texts*, in Proceedings of the 2nd Romanian Workshop for Linguistic Tools and Resources Volume, 3 Nov. 2006, Bucharest, Romania.

Steinberger, R., Pouliquen, B., Hagman, J. (2002). *Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc*, Springer-Verlag.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 2142--2147. Genoa, Italy, 24-26 May 2006.

Strzalkowski T. (1995). *Natural Language information retrieval*, IP&M, 31(3), pp. 397--417.

Todiraşcu, A., Gledhill, C., Ştefănescu, D. (2009). *Extracting Collocations in Contexts*, in Human Language Technology. Challenges of the Information Society, LNCS Series, Springer, Vol. 5603/2009, pp. 336--349. ISBN 978-3-642-04234-8.

Tufiş D., Ion R., Ceauşu A., Ştefănescu D. (2008). *RACAI's Linguistic Web Services*, in Proceedings of the 6th Language Resources and Evaluation Conference (LREC), Marrakech, Morocco, pp. 28--30.

Tufiş D., Irimia E. (2006). *RoCo_News - A Hand Validated Journalistic Corpus of Romanian*, in Proceedings of the 5th LREC Conference, Genoa, Italy, pp. 869--872.

Velardi, P., Navigli, R., D'Amadio, P. (2008). *Mining the Web to Create Specialized Glossaries*, IEEE Intelligent Systems, 23(5), IEEE Press, pp. 18--25.

Weller, M., Gojun, A., Heid, U., Daille, B., Harastaniv, R. (2011). *Simple methods for dealing with term variation and term alignment*. In Proceedings of TIA 2011: the 9th International Conference on Terminology and Artificial Intelligence, November 8-10, Paris, France.

Wong, W., Liu, W. and Bennamoun, M. (2007). *Determining Termhood for Learning Domain Ontologies in a Probabilistic Framework*. In 6th Australasian Conference on Data Mining (AusDM).