Turning a term extractor into a new domain: first experiences

Anna Joan, Jorge Vivaldi, Mercè Lorente
Institute for Applied Linguistics,
Universitat Pompeu Fabra
Plaça de la Mercè, 10-12,
08002 Barcelona, Spain

E-mail: {anna.joan, jorge.vivaldi, merce.lorente}@upf.edu

Abstract

Computational terminology has notably evolved since the advent of computers. Regarding the extraction of terms in particular, a large number of resources has been developed: from very general tools to other much more specific acquisition methodologies. Such acquisition methodologies range from using simple linguistic patterns or frequency counting methods to using much more evolved strategies combining morphological, syntactical, semantical and contextual information. Researchers usually develop a term extractor to be applied to a given domain and, in some cases, some testing about the tool performance is also done. Afterwards, such tools may also be applied to other domains, though frequently no additional test is made in such cases.
Usually, the application of a given tool to other domain does not require any tuning. Recently, some tools using semantic resources have been developed. In such cases, either a domain-specific or a generic resource may be used. In the latter case, some tuning may be necessary in order to adapt the tool to a new domain. In this paper, we present the task started in order to adapt YATE, a term extractor that uses a generic resource as EWN and that is already developed for the medical domain, into the economic one.

1. Introduction

Since the 80s, multiple efforts have been focused on term extraction. Many techniques have been developed, some of them using general purpose tools while others using specific tools (term extractors, –TE–). Often, such TEs take profit of techniques and resources developed for a specific domain and language. Moreover, such TEs are developed to fulfil some particular needs like glossary compilation, translation or NLP purposes (Information Retrieval, ontology/conceptual map generation, etc.).
Usually, researchers develop a TE for a given domain but they seldom try to adapt and/or test the tool (and the specific resources, if any) to a different domain. In this paper, we discuss the problem of turning a TE developed for a given domain into a different one.
The paper will be organised as follows: Section 2 briefly introduces the TE's state-of-the-art and Section 3 presents YATE, the TE system developed for the medical domain that has been turned into Economics. Section 4 describes the adaptation task in some detail. Section describes the evaluation of the TE in the new domain. Finally, Section 6 presents some conclusions as well as some suggestions about future work to be done to ease the adaptation procedure.

2. Term extraction: state-of-the-art

In the last two decades many TEs have been developed. Most of them follow either a linguistic or a statistical approach (see Cabré *et al*. 2001 and Kageura *et al*. 1996). However, hybrid approaches have also been adopted to overcome the limitations of the previous ones. Due to the difficult of assuring that a given lexical unit belongs to a specialised domain, the result obtained by all the TEs are "term candidates" (TC) instead of just "terms". This means that such tools are just support systems offering proposals to terminologist, who must decide how to proceed.
Usually, terms are described by a sequence of contiguous parts-of-speech[1] whose size ranges from a single word to a complete noun phrase. This fact is at the base of most of the approaches to term extraction, which turns them into being language dependant. TERMS (Justeson *et al*., 1995) is a term extraction tool usually considered as the prototypical system in using this approach. Some variation is found in TERMINO (Plante *et al*., 1989), (Heid *et al*., 1996) and LEXTER (Bourigault, 1994). This strategy usually results in a huge number of TC that have to be manually checked. An exception of this statement is FASTR (Jacquemin, 2001), because although it is a fully linguistic method, it obtains very good results mainly due to the fact that it is a term variant detector instead of a true term acquisition tool. A quite different system is described in (Ananiadou, 1994); it makes profit from the fact that medical terminology relies heavily on Greek and Latin neoclassic elements, even for the creation of new terms. This system has a high precision but a limited recall; also, it may be applied to a limited number of domains.
The second family of tools are those based in statistics. In this case, they are not used alone but in combination with some linguistic technique to improve their results. Linguistic knowledge can be used *a priori* as in ACABIT (Daille, 1994), ANA (Enguehard, 1993) and TermoStat (Drouin, 2002), or *a posteriori* as in (Smadja, 1993).
Usually, the results obtained using any of the above mentioned approaches are poor if they are applied in a strict way. As mentioned, both systems must make use

[1] For example: ((A|N)+|((A|N)*(NP)?)(A|N))*N for English, N(A|P(D)?N) for French and N(A+|P(D)?NA+) for Spanish and Catalan; where N=noun, A=adjective, P=preposition and D=determiner.

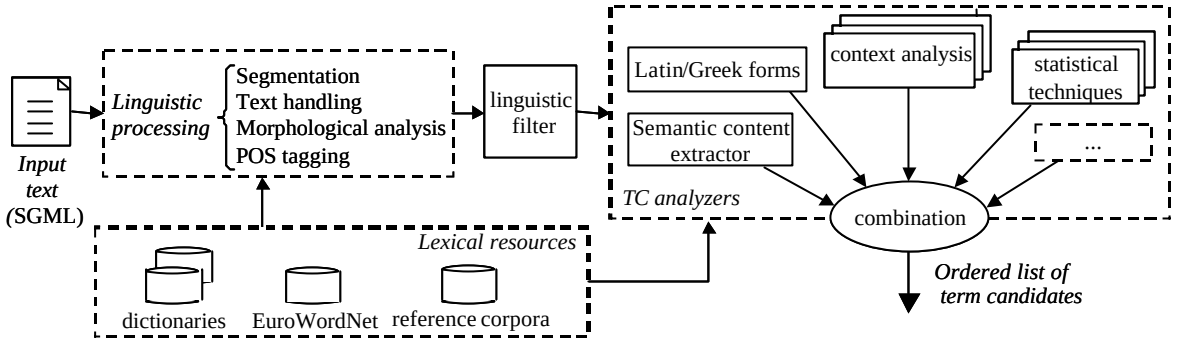


Figure 1. Architecture of YATE

of a component, even if minimal, of the other approach in order to reach some effectiveness. In spite of this combination, the above mentioned tools heavily rely on one of the approaches. Recently developed methods tend to a more equilibrate use of linguistic and statistical data; therefore, they are usually seen as hybrid systems. Another main characteristic is that they use some kind of semantic information.

TRUCKS (Maynard, 1999) is a hybrid tool that consecutively applies a number of statistical measures, starting from pattern based TC selection. It is interesting to note that such measures take profit of the context to calculate the termhood of a TC and one of them relies in the consultation to the semantic network of the UMLS [2].

Some of the above mentioned TE systems have been conceived without any restriction in its application domain, while others are restricted by design (Ananiadou 1994 and TRUCKS[3]). Anyway, they have been mostly tested for just a single domain and no attempt has been made to check in depth their behaviour in other domains.

3. YATE: A term extractor[4]

In this paper, we have worked on the adaptation of YATE (see Vivaldi 2001a for details), a term extraction tool whose main characteristics are: a) it uses a combination of several term extraction techniques and b) it uses EWN[5], a general purpose lexico-semantic ontology as a primary resource.

YATE was first designed to obtain all terms (from the following set of syntactically filtered candidates: <noun>, <noun-adjective> and <noun-preposition-noun>) found in medicine specialised texts. YATE is a hybrid TE system combining the results obtained by a set of term analyzers (TA) described briefly as follows (see Figure 1):

a) Domain coefficient (DC): uses the EWN ontology to sort candidates.
b) Context (CFp): evaluates each candidate using other candidates from its context.
c) Classic forms: decomposes lexical units into formants, taking into account terms formal characteristics in some domains.
d) Collocational method: evaluates multiword candidates according to its association score.

The results from this set of heterogeneous methods are combined using voting and boosting (see Vivaldi et al., 2001b and 2002). See Vivaldi et al. (2007) for a discussion about TE evaluation and the methodology followed to evaluate this tool.

It is interesting to note that the TAs a) and b) require the use of EWN, and also that the combination of different TA gives priority to a) and c).

In using this kind of resources it is necessary to determine whether a certain word belongs or not to a given domain. For such purpose, we define the notion of domain border (DB). A synset may be defined as a DB when all its hyponyms also belong to such domain. For example, in medicine, the synset "08577911n" (physiological state) is a DB because all the diseases registered in EWN are its hyponyms. See Vivaldi *et al*. (2002) for details.

Obviously, this kind of data is domain dependant; therefore, the use of this TE in domains other than medicine requires the tuning of a number of configuration files of YATE[6].

4. Adaptation

In this section we will briefly present how the adaptation will benefit both nouns and adjectives that are part of TCs (section 4.1). Also we will show the procedure to be followed in this adaptation procedure (section 4.2).

[2] UMLS: Unified Medical Language System (http://www.nlm.nih.gov/research/umls/).

[3] TRUCKS usage is limited to medical domain. The statistical measure NC-value (Frantzi, 1997) is a basic part of this TE and it has been also implemented in a number of other TE for a number of different domains.

[4] http://igraine.upf.es:4000/main

[5] EWN (http://www.illc.uva.nl/EuroWordNet/) is a multilingual extension of WordNet, a lexico-semantic ontology developed at Princeton University. The basic semantic unit is the synset (synonymy set), grouping together several words that can be considered synonyms in some contexts. Synsets are linked by means of semantic labels (hyperonym, hyponym, meronym, etc.). Due to polysemy, lexical entries can be attached to several synsets.

[6] The usage of EWN has the advantage of being a general purpose resource that can be adapted to different domains (the price to pay is its limited coverage). Specialised ontologies are more exhaustive but limited to just one domain; moving to another domain implies finding a new resource and building an interface specific to such resource.

4.1 Advantages

As mentioned above (see section 3), YATE requires to know if a given word belongs or not to the domain of interest. This means, for example, that if a given noun is not a hyponym of a DB it will not be well ranked for the TA that uses the DC. Taking into consideration that the output of YATE is a ranked list of units, the inclusion of a noun in the EWN hierarchy will imply the climbing of such noun in the rank obtained from the DC method. This may also indirectly benefit other TC having such noun in their context.

In the case of adjectives, the situation is similar but the benefit may affect more that one term. Consider for example the Spanish adjective *arancelario* ("tariff" used as an adjective), that may be used in several terms being relevant for the economy domain as: *código arancelario* ("tariff code"), *barrera arancelario* ("tariff barrier"), *sistema arancelario* ("tariff system"), *reforma arancelaria* ("tariff reform"), *deficit arancelario* ("tariff gap"), etc. Therefore, the inclusion of this adjective will involve detection improvement in the case of several terms in the domain.

4.2 Procedure

The adaptation of YATE into a new domain may be roughly divided in two main stages (Vivaldi, 2006): the first one regards the resources and utility files used in the tool and the second one the data required to evaluate it. Nevertheless, these working tasks are considered to be interrelated, as the results of each one are used to improve the others. The first main stage may be also divided in two sub stages: the first one closely related with the tools itself (some configuration files) and the last one, involving the EuroWordNet ontology enlargement, which is a main resource for YATE (see section 3). The second main stage focuses on evaluating the adaptation made.

YATE requires three utility files to work: a) domain borders, b) properties relevant to the domain and c) combination of domain borders and properties that are also relevant to the domain. A domain border is a synset in EWN from which we may establish that all hyponyms pertain to the domain under study. This file is essential for the extractor as almost every TA uses it directly, or indirectly. The property file collects all the synsets that receive a value through the qualifying adjectives and are relevant to the domain. The last file assembles the couple domain border – property that are relevant to the domain. The last two files allow to detect terms belonging to the <noun-adjective(qualifying)> pattern.

At first, we work just in the definition of the domain borders in Economics. To have some clues on where to start from in the subborder establishment, a previous prospecting on the data of the available corpora is required to find the most productive units in the domain we want to adapt YATE to. For such purpose we use a concordancer to query the IULA LSP corpus[7].

Such units are used to determine the DB in the EWN nominal hierarchy. For this task, we proceed in the following way: a) finding such units in EWN; b) if there is polysemy, locating the sense/s that apply to Economics; c) if the synset is a potential DB, analyzing the hypernymy chain to look for a more general DB (or, covering a higher number of synsets) and d) repeating the process with the following productive word. In doing this task we may identify the following situations and actions:

a) The noun is already included in EWN.
b) The same of the above but it is included just in English section of EWN: include the word in the Spanish section.
c) The word is not included in the EWN hierarchy: add the word in the nominal hierarchy of EWN.

Although the above situation reflects the procedure followed just for nouns, there is similar one for adjectives (relational/qualifying). In both cases, reference specialized dictionaries and experts are consulted, to verify the adequacy of the new synset in EWN. Domains like Medicine and Economics are different in the units used as specialized in their discourse. Medicine units are found agglomerated in specific points of the hierarchy and we easily find verticality in the hierarchy once a domain border is established, while domain borders in Economics, a field with a discourse nearer to general language, are set scattered in different points of the hierarchy so that the feeling is that horizontality is predominant in this field.

As for the second main stage, or the evaluation step, a text in Microeconomics and Spanish language was chosen[8]. Three economy experts were asked to underline the (group of) words that they considered to be specialized in the domain. The words chosen were placed in a database, classified according their syntactic pattern.

5. Evaluation

Virtually all TE systems have their origins in information retrieval or Linguistics. The former focuses its evaluation measures in precision and recall measures while the latter is based on noise and silence figures. Both perspectives give basically the same information but in a different way and has been largely described on the literature.

YATE belongs to the class of TE named rankers because it ranks the term candidates according to their "termhood". This figure has been defined in Kageura et al. (1996) as "the degree that a linguistic unit is related to domain-specific concepts"; which seems suitable for this task.

To evaluate YATE in the economics domain, we followed the same procedure used in the medical domain[9]: a) to use the glass box evaluation model and b) to use an intrinsic method. The first decision means that we evaluate for both TAs and combination method. This is important because the method DC has influence in the final result but also in the performance of other methods. The second decision implies that we need to obtain the list of terms included in the test document (see section 4) in order to

[7] It refers to the Technical Corpus hosted in the Institute for Applied Linguistics (see Badia et al, 1998). It may be consulted using the corpus browser *bwanaNet* (http://bwananet.iula.upf.edu/).

[8] The chosen text is *Acerca de la confianza en el dinero* (J. Esteban), a chapter of a university textbook and 45,984 word-sized.

[9] The results achieved by applying YATE to the medicine domain are much better. We evaluated the tool with a specialised document of about 100K words (a collection of medical reports). For a recall of 30%, the precision was the following one: nouns 95% and noun-adjective 75%. See Vivaldi et al. (2007) for details.

calculate the precision versus recall curves for each syntactic pattern.

But there is a previous consideration to be done: there are discrepancies about what a term must be considered to be; such disagreement can be found among terminologists and specialists but also among specialists. This fact, initially found in the medicine domain, is now confirmed in economics as shown in Table 1: the full agreement is only 31.6 % for all the patterns but a bit higher (36.8 %) considering just the patterns processed by YATE.

Pattern	1 evaluator	2 evaluators	3 evaluators	Total
N	91	67	92	250
NJ	168	154	200	522
NPN	110	87	102	299
Others	216	143	85	444
Total	585	451	479	1515
	38.6%	29.8%	31.6%	
Total (N+NJ+ NPN)	369	308	394	1071
	34,5%	28,8%	36,8%	

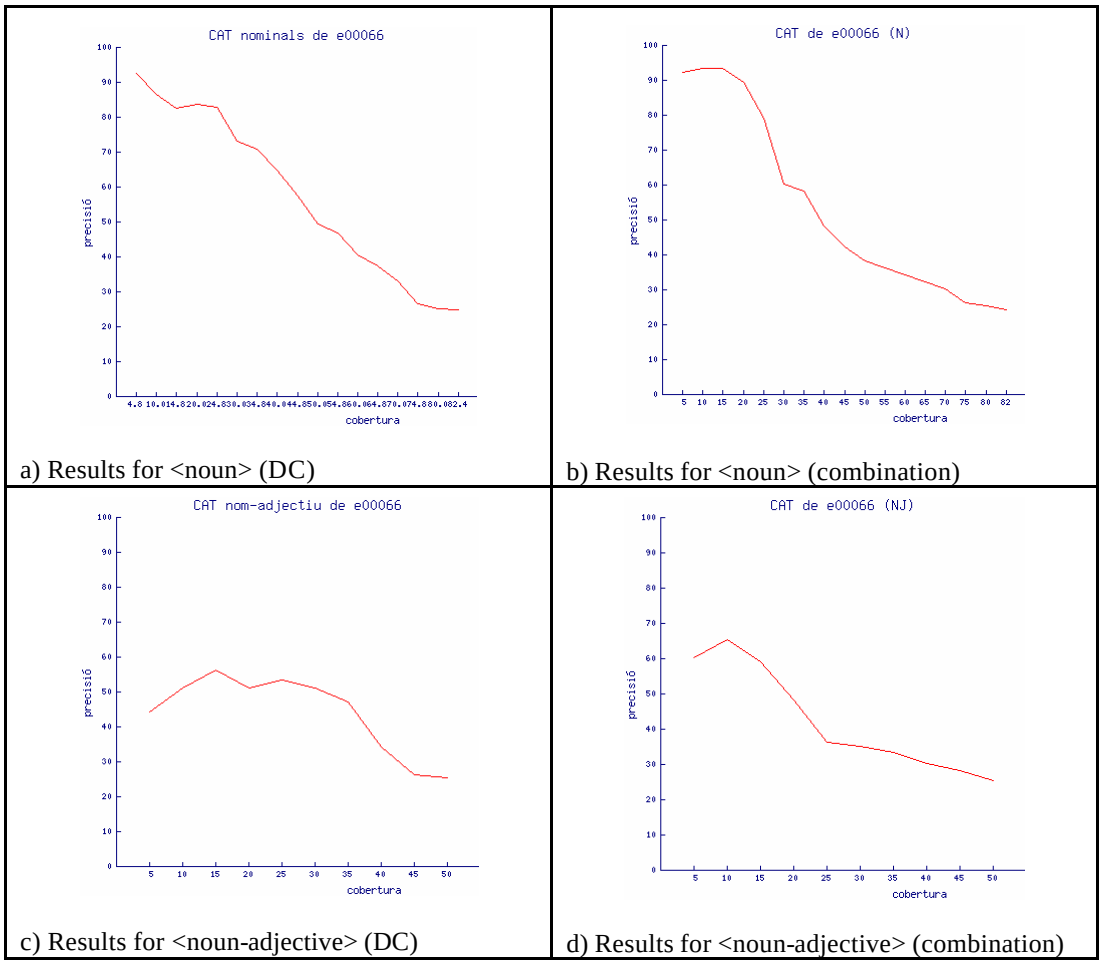
Table 1. Terms chosen by the specialists.

The disagreement among evaluators is common to other NLP activities that also require manual validation like word sense disambiguation, discourse analysis or POS tagging.

In our case, the disagreement may also be found in that the "perception" of units as being specialized is determined by the target of the detection as well as the different approaches from which the experts have come to master the knowledge of the domain.

Other reason of disagreement would be that manual tasks are open to mistakes: consider the terms "net price", "competitive price" or "optimal price"; in spite of their clear terminological nature, each of them has only been tagged by one specialist. Another example is the term *comercio* ("commerce"), it has been only tagged in sequences like *comercio exterior* ("foreign commerce") but never in isolation. Finally, specialists tagged a sequence like *área de comercio exterior* ("foreign market area") but it should be considered as a phraseological unit, only the subsequence *comercio exterior* is a term.

Figure 2 shows the results obtained for patterns <noun> and <noun-adjective> only using the DC method (a and c) and a combination method (b and d). By analyzing such curves, we observed that the main hypothesis of YATE (combination methods perform better that isolated methods) is also valid in this domain. Moreover, the behaviour of the pattern <noun-adjective> is lower than medicine, mainly due to the lack of some adjectives in EWN and the lack of refinement of the related configuration files in YATE.



a) Results for <noun> (DC)

b) Results for <noun> (combination)

c) Results for <noun-adjective> (DC)

d) Results for <noun-adjective> (combination)

Figure 2. Results obtained by YATE

When analyzing the results more in depth it is possible to observe several phenomena:

a) Some terminological units have not been tagged by any specialists (*economista* –economist-, *comercio* -commerce-) even if their contexts are relevant.
b) Some nouns (*sobreinversión* -overinvestment-, *duopolio* -duopole-, etc.) and adjectives (*oligopolístico* -oligopolistic-, *paretiano* -paretian-, etc.) are still missing in EWN.
c) Terms including specifiers are found but not processed by YATE as for example *crecimiento de la economía* (economy grow) or *elasticidad de la demanda* (demand elasticity). These and similar patterns were not previously taken into account due to their unproductivity in medicine discourse.
d) Words that are proposed by YATE due to POS tagging errors. See for example: *general* -general- like a noun instead of adjective or *debe* -debit- as a noun instead of verb.
e) As foreseen, the TA using Latin Greek formants is not efficient in economics as in medicine.

6. Conclusions

This paper shows the process followed to adapt a TE developed for a domain (Medicine) to a new domain (Economics). It is an iterative process and includes a performance evaluation using precision and recall measures. We showed that, by result observation, it is easy to obtain the different aspects needed to be reconsidered on an efficient adaptation procedure.

The process of adaptation of a TE into a new different domain is an iterative process. After an initial step, it is necessary to proceed in the sequence of enlarging EWN, refining the configuration files and analyzing the results. Moreover, the TE itself needs refining (to allow for more morphosyntactic patterns to be detected and for each specific domain term features to be considered) as well as improving of some of its TA.

As a future task, we plan to complete, as in depth as necessary, the adaptation process enlarging EWN as well as to process more documents in the domain and evaluate the final result. Also it will be necessary to improve the TE, including the processing of some new patterns, and to explore the possibility to include a method exploiting common prefixes as a replacement/enlargement of the Latin/Greek TA module.

7. Acknowledgements

This paper has been written within the research projects RICOTERM2 (HUM2004-05658-C02-01) (Lorente, 2006) and RICOTERM3 (HUM2007-65966-C02-00), financially supported by the Spanish government.

8. References

Ananiadou, S. (1994) "A methodology for automatic term recognition". *In Proceedings of the Fifteenth International Conference on Computational Linguistics* (Coling'94). Kyoto, pp. 1034-1038.

Badia, T.; Pujol, M.; Tuells, A.; Vivaldi, J.; de Yzaguirre, L. y M. T. Cabré (1998) IULA's LSP Multilingual Corpus: compilation and processing. *In Proceedings of* the *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Bourigault, D. (1994). *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes*. Ph.D. thesis. École des Hautes Études en Sciences Sociales, Paris (France).

Cabré, M. T., R. Estopà & J. Vivaldi, (2001). Automatic Term Detection: A Review Of Current Systems. In Bourigault, D., C. Jacquemin and MC. L'Homme (eds) *Recent Advances in Computational Terminology*. Chp 3. Amsterdam: John Benjamins

Drouin, P. 2002. Acquisition automatique des termes: l'utilisation des pivots lexicaux spécialisés. Ph.D. Thesis. Montreal (Canada): Université de Montréal.

Frantzi, K. T. (1997). Incorporating context information for extraction of terms *Proceeding of ACL/EACL-97*. Madrid, pp. 501-503.

Heid U., S. Jau, K. Kruger and A. Hohmann (1996). Term extraction with Standard tools for corpus exploration. Exploration from German. Proceedings of TKE '96, (Frankfurt: Indeks), pp. 139-150.

Jacquemin C. (2001). *Spotting and discovering terms through natural language processing*. Cambridge (Mass.) MIT Press.

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

Kageura K. and Umino B. (1996). Methods for Automatic Term Recognition: A Review. *Terminology* 3(2), pp. 259-289.

Lorente, M. (2006) Expansió de consultes multilingüe per a la recuperació d'informació en economia. Juan, M.; Amengual, M.; Salazar. J. (ed.) *Lingüística aplicada en la sociedad de la comunicación y la información*. Palma de Mallorca. ISBN 84-7632-962-8 [Conference and papers of the XXIII Congreso Internacional de AESLA (2005)].

Maynard, D. (1999). Term Recognition Using Combined Knowledge Sources. Ph.D. Thesis. Manchester Metropolitan University. Manchester (UK).

Plante, P and Dumas L. (1998) "Le dépoulliment terminologique assisté per ordinateur". Terminogramme, 46, pp. 24-28.

Smadja, F. (1993). Extracting collocations from text: Xtract. Computational Linguistics. Vol. 19 (1), pp. 143-177.

Vivaldi, J. (2006). *Sistema de extracción de candidatos a término YATE: manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada. Sèrie Informes, 43.

Vivaldi, J. (2001). Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD thesis. Universitat Politècnica de Catalunya. Barcelona, Spain.

Vivaldi J. & Rodríguez H (2007). Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13:2, pp. 225–248.

Vivaldi, J. & Rodríguez, H, (2002). Medical Term Extraction using the EWN ontology. In *Proceedings of Terminolgy and Knowledge Engineering (TKE2002)*.

Vivaldi, J. & Rodríguez, H, (2001). Improving term extraction by combining different techniques. *Terminology*. John Benjamins Publishing Co 7:1, pp. 31-47.