Improving Term Candidate Validation Using Ranking Metrics

Article · June 2013

CITATION
1

READS
205

2 authors:

Mercè Vázquez
Universitat Oberta de Catalunya
20 PUBLICATIONS **92** CITATIONS
SEE PROFILE

Antoni Oliver
Universitat Oberta de Catalunya
113 PUBLICATIONS **463** CITATIONS
SEE PROFILE

AWERProcedia Information Technology & Computer Science

Vol 03 (2013) 1348-1359
3rd World Conference on Information Technology (WCIT-2012)

Improving Term Candidate Validation Using Ranking Metrics

Mercè Vàzquez *, Information and Communication Sciences Department, Universitat Oberta de Catalunya, Rambla Poblenou 156, 08018, Barcelona, Spain.
Antoni Oliver, Arts and Humanities Department, Universitat Oberta de Catalunya, Avinguda Tibidabo 39-43, 08035, Barcelona, Spain.

Suggested Citation:
Vàzquez, M. & Oliver, A. Improving Term Candidate Validation Using Ranking Metrics, *AWERProcedia Information Technology & Computer Science*. [Online]. 2013, 3, pp 1348-1359. Available from: http://www.world-education-center.org/index.php/P-ITCS *Proceedings of 3rd World Conference on Information Technology* (WCIT-2012), 14-16 November 2012, University of Barcelon, Barcelona, Spain.

Received 19 January, 2013; revised 11 June, 2013; accepted 17 September, 2013.
Selection and peer review under responsibility of Prof. Dr. Hafize Keser.

Abstract
At times it is difficult to automatically identify the most representative terms in a specialized corpus and to validate them as correct due to the similarity of words and terms. In order to identify the most representative terms in a corpus that can be easily adapted to any language or terminology extraction tool, we explore the combination of token slot extraction and ranking metrics to select term candidates with a high likelihood of being terminological units. This paper presents the results we have identified using four statistical measures. We observe high term detection in English corpora (a precision of 76.92% and a recall of 79.09%) and Spanish corpora (a precision of 60% and a recall of 70.48%) using token slot detection together with four ranking metrics: Dice, True Mutual Information, T-score and Log-likelihood. In conclusion, token slot detection extracts terminological patterns in term candidates to reduce lists of candidates, and ranking metrics improve results and reduce the number to be evaluated manually. We will evaluate the algorithm's performance in other domains and for other user profiles and needs.

Keywords: Term candidate validation, ranking metrics, term extraction, token slot detection;

* ADDRESS FOR CORRESPONDENCE: **Mercè Vàzquez,** Information and Communication Sciences Department, Universitat Oberta de Catalunya, Rambla Poblenou 156, 08018, Barcelona, Spain, *E-mail address*: mvazquezga@uoc.edu

1. Introduction

Terminologies are becoming increasingly important in everyday life as technology and science continue to grow at an accelerating rate. Since the 1990s, an increasing amount of terminology research has been devoted to facilitating and augmenting terminology-related tasks by using computers and computational methods. One focus for this research is Automatic Term Extraction (ATE), term extraction specifically done using computational methods [1]. The notion of *term* in this context can be defined as a "linguistic representation of concepts" [2]. In specialized domains, terms are used to identify concepts in order to provide up-to-date terminological material, aid the work of writing, editing, translation, and lexicography and terminography professionals or promote multilingual work in general. Likewise, collecting terms improves the tasks of collating, classifying and cataloguing information, and thesaurus compilation, and makes (monolingual and bilingual) information retrieval much easier [3], [4], [5].

Research on automatic terminology extraction uses either linguistic specifications, statistical approaches or hybrid approaches. Concerning the former, Bourigault [6] has proposed a program which can extract from a corpus sequences of lexical units whose morphosyntax characterizes technical noun phrases automatically. Daille [7] and Jacquemin [8] propose a morphological and syntactical analysis using dependency analysis. Likewise, Pazienza [9] discusses which is the best term extraction process based on linguistic resources. The final list of sequences is given to a terminologist to be checked. For the latter, several works [10], [11], [12], [13] have shown that statistical scores are useful to extract collocations from corpora. The main problem with one or the other approach is the "noise": indeed, morphosyntactic criteria are not sufficient to isolate terms, and collocations extracted as a result of statistical methods belong to various types of associations: functional, semantic, thematic or others which are uncharacterizable. Furthermore, term properties have no formal rules to use in an automatic term extraction process [14], and also term variation adds difficulty to automatic term identification, because one term can be related to multiple concepts (polysemy), but it has to be considered as a part of term mining [15], [16], [17], [18]. A recent study using hybrid approaches [19], that is, part-of-speech tagging and relative frequency, demonstrates the complexity of identifying terms from a corpus: errors in text preprocessing steps, noise stems from corpora, lack of correspondence between candidates or term variants affects the extraction process negatively. Thus, there is still the need for a method which can extract term candidates from specialized corpora while avoiding the processing drawbacks as described above [20], [21], [22] being useful to any language (especially minority languages) [23], [24] and facilitating the term candidate validation task.

To improve these approaches to term candidate validation, we explored the performance of an algorithm based on statistical methods to extract terms from specialized domains and we investigated how the overall performance of multi-word term extraction from a specialized corpus can be significantly improved by enriching reference term lists with automatically extracted domain-specific terminological tokens. To do so, we combined a recursive use of token slot extraction and ranking metrics as a basis for term extraction. An appropriate ranking metric correlates to the precision of a term candidate: using an effective term ranking metric makes it possible to select a set of term candidates which when processed will result in a higher number of approved terms compared to selecting the set of term candidates to be processed randomly or using a poor ranking metric [25].

This paper describes experimental results obtained applying the statistical algorithm in several specialized domains (Telecommunications and Economics) and languages (Spanish and English) and shows how this approach can be applied to validate term candidates in the languages used while overcoming the processing problems.

This paper is structured as follows: in the next section we present the algorithm and all the resources and tools that were used in the experiment. The results and discussion are described in detail in Section 3. The paper is concluded with some final remarks and ideas for future work.

2. Resources and Tools

Automatic term detection from specialized corpora is a complex task considering the similarity of words and terms and computational complexity needed to distinguish them, so the term candidates list extracted to be validated manually can be extensive. In order to improve the term candidate validation task, we propose a statistical based algorithm which selects term candidates from a specialized corpus with a high likelihood of being terminological units. To do so, the algorithm enriches reference term lists (gold standard) with automatically extracted domain-specific terminological tokens using a recursive process which combines token slot extraction and ranking metrics. Thus, the greater the number of reference terms, the greater the number of candidates that can be filtered and selected as terms. We consider a token as an instance of a sequence of characters in a given document that are grouped together as a useful semantic unit for processing. We believe that using tokens from reference terms (terminological token patterns) helps to select the most relevant term candidates. Furthermore, this algorithm can be applied easily to any language (which is especially useful for minority languages), and to different users' needs or in tools which use terminological resources (computer-assisted translation systems, machine translation or terminological tools).

The proposed approach for extracting multi-word terms from specialized domains is composed of three main steps: (i) token slot extraction, where we extract bigrams from corpora, (ii) ranking metrics, where we rank bigrams using the True Mutual Information, T-score, Log-likelihood and Dice scores, (iii) reference term list enrichment, where new manually-selected terms are used to enrich the reference term list and improve the token slot extraction. In the following subsections, we cover the three steps in more detail.

2.1. Token slot extraction

The terminology extraction process starts when a list of bigram term candidates has been extracted from specialized corpora, and the results filtered with a list of stop words (functional or connective words that are assumed to have no information content) and ordered by frequency. To do this, we use open-source software, the Ngram Statistics Package [26]. The algorithm proposed then assigns each term candidate the status of being a term if each token of the term candidate is found in the reference term list for each domain (gold standard). A term candidate such as "digital processing" in the Telecommunications domain contains two token slots in which slot 1 is filled by "digital" and slot 2 by "processing". The algorithm selects this term candidate from the list of term candidates if one or more such slots can be filled by terminological tokens from the reference terms: e.g., "digital networks" and "signal processing" [27]. If a list of reference terms is not available, there is the possibility of selecting candidates manually. Terms selected will be used by the algorithm as reference terms to prepare token slot extraction recursively. Table 1 shows an example of token slot extraction.

Table 1. Example of token slot extraction

Reference term	Term candidates (slot 1)	Term candidates (slot 2)	Terms selected (slot 1)	Terms selected (slot 2)
Telematic terminal	Telematic access Telematic interworking Telematic user	Tdma terminal remote terminal Vsat terminal	Telematic access Telematic interworking	Tdma terminal remote terminal

2.2. Ranking metrics

Following the token slot extraction step, the algorithm ranks term candidates filtered by the reference term list using four different ranking metrics that are widely applied in terminology extraction: True Mutual Information, T-score, Log-likelihood and Dice [28], [29]. The ranking order produced by these metrics is then compared in terms of accumulated precision. Figure 1 shows the proposed ranking metrics algorithm.

Contingency table

	$V=v$	$V\neq v$
$U=u$	$O11$	$O12$
$U\neq u$	$O21$	$O22$

U: first word of the bigram *V*: second word of the bigram
O11: #compound words with *U* and *V*
O12: #compound words with *U* but without *V*
O21: #compound words with V but without U
O22: #compound words without U and without V

$$MI = \log \frac{O11}{E11} \qquad T - score = \frac{O11 - E11}{\sqrt{O11}} \qquad Dice = \frac{2O11}{R1 + C1}$$

*E*11 (expected co-occurrence frequency) is small

R1=O11+O12 C1=O11+O21
R2=O21+O22 C2=O12+O22

$$N=O11+O12+O21+O22=R1+R2=C1+C2$$

$$LLR = -2\log\left\{\frac{L(O11,C1,r)*L(O12,C2,r)}{L(O11,C1,r1)*L(O12,C2,r2)}\right\}$$

$$L(k,n,r) = r^{k}x1-r^{n-k}$$

$$r=\frac{R1}{N} \quad r1=\frac{O11}{C1} \quad r2=\frac{O12}{C2}$$

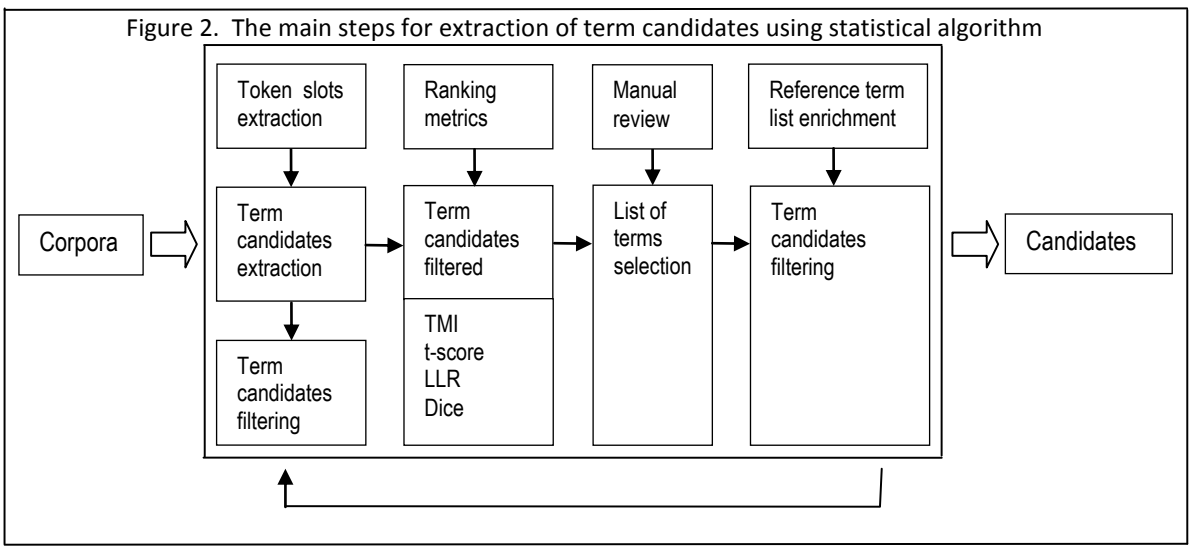
Figure 1. Ranking metrics equations

2.3. Reference terms list enrichment

Following the ranking metrics step, a list of term candidates is produced for manual review by an expert. After this point, those candidates selected as terms are used by the algorithm to filter term candidates which have not been taken into account in the first token slot extraction step, and so on, recursively.

In the last step of the terminology extraction process the reference terms lists are enriched to filter out token slots from the complete list of term candidates. This step is carried out recursively. Figure 2 shows the main steps for extraction of term candidates using the statistical algorithm. Figure 3 shows the statistical algorithm.

Figure 2. The main steps for extraction of term candidates using statistical algorithm



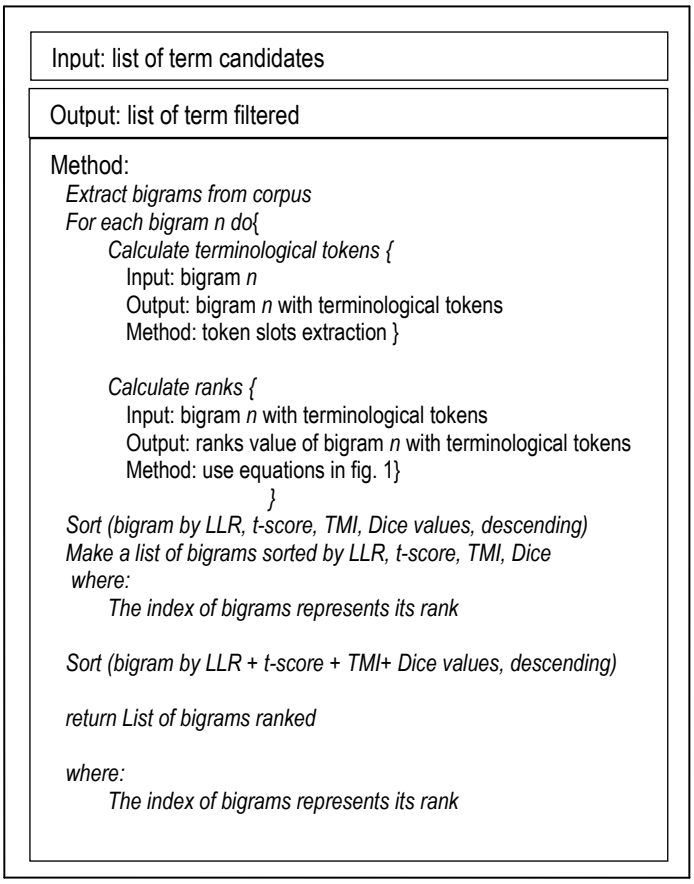


Figure 3. The statistical filter algorithm

3. Results and Discussion

We have tested the proposed statistical algorithm on two specialized corpora. In this section we provide the results of our experiments.

3.1. The corpus

The algorithm proposed has been explored in the Telecommunications and Economics domains. A specialized Telecommunications corpus (Crater) [30] was used to extract the list of term candidates in English. The corpus is a trilingual (English, French and Spanish) parallel aligned corpus. As the domain of the three component corpora is telecommunications, they are a particularly good resource for studying automated terminology extraction.

A specialized Economics corpus [31] was used to extract the list of term candidates in Spanish. The corpus was compiled by the Institute for Applied Linguistics (IULA), a research and graduate training center at Universitat Pompeu Fabra. The corpus contains 41,385 words.

Reference term list

For the Telecommunications domain, the reference terms list was made up of terms from the *Diccionari de telecomunicacions* published by Universitat Politècnica de Catalunya in 2007 [32] and available in English, Spanish and Catalan. It contains 1,000 bigram terms. For the Economics domain, the reference terms list was made up of terms from the Institute for Applied Linguistics (IULA) corpus available in English, Spanish and Catalan. It contains 518 bigram terms.

3.2. Results and evaluation

Evaluation of automatic terminology extraction is always a complex task, because there are no specific standards to evaluate and compare different approaches. However, most of the approaches have used reference terms list and validation [9]. We combine these two approaches to evaluate our results: term reference lists are used to filter term candidates (token slot extraction) and manual term candidates review by an expert are done to enrich term reference list and filter recursively term candidates.

Indeed, we assess the performance of the algorithm proposed in terms of precision and recall. Precision measures the correctness of the lexical units that are suggested as terms, usually measured as the ratio of correct ("true positives") and all suggested units ("true positives" and "false positives"). Recall denotes the degree to which concepts in a document are recognized, usually measured by the ratio of the correctly recognized terms ("true positives") and all domain-relevant terms occurring in a given document ("true positives" and "false negatives"). The overall performance is measured by a single score, called the F-measure [33]:

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

The results of our approach are as follows. First, term candidates were extracted from specialized corpora, filtered by a list of stop words and ordered by frequency. We obtained 92,428 term candidates (Telecommunications corpora) and 2,214 term candidates (Economics corpora), respectively. Then, during the token slot extraction step, the algorithm selected those term candidates where one or more slots could be filled by terminological tokens from the reference terms. Following token slot extraction, we obtained 3,385 term candidates in the Telecommunications domain and 192 term candidates in the Economics domain. To do so, for the Telecommunications domain we used a list of 500 reference terms (randomly selected from a 1,000 bigram terms list), 400 of which were on the term candidates list. Regarding the Economics domain, we used a list of 300 reference terms (randomly selected from 518 bigram terms list), 110 of which were on the term candidates list.

Second, the term candidates obtained were ranked by the algorithm using four different ranking metrics: True Mutual Information, T-score, Log-likelihood and Dice. The ranking order produced by these metrics was then compared using accumulated precision. Figure 4 (a) shows how 400 reference terms from the Telecommunications domain are distributed by these four ranking metrics and figure 4 (b) shows how 110 reference terms from the Economics domain are distributed by these four ranking metrics.

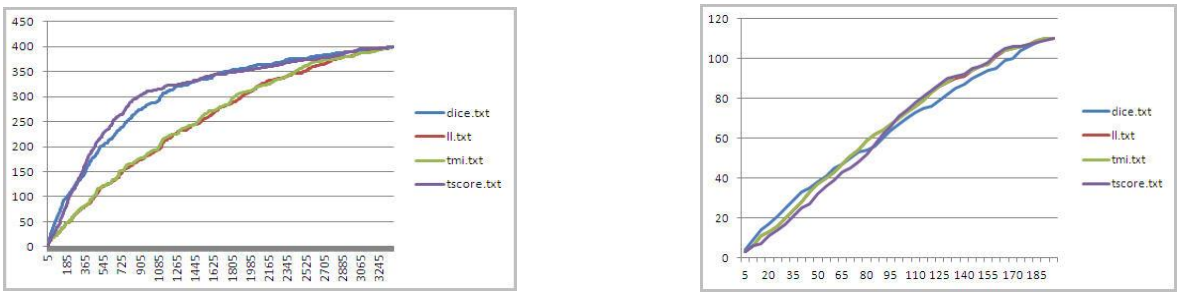


Figure 4. (a) 400 reference terms list distribution (Telec); (b) 110 reference terms list distribution (Economics)

Using these ranking metrics in the Telecommunications corpora most of the 400 reference terms are placed between positions 2,000 (Dice) and 2,500 (T-score) on the list of 3,385 term candidates. In consequence, by validating only the first 2,500 term candidates, instead of all 3,385, we obtained 375 terms. Table 2 shows how reference terms from the Telecommunications domain are distributed by these four ranking metrics.

Table 2. Number of reference terms list distribution (Telecommunications)

Method	Top 500	Top 1,500	Top 2,000	Top 2,500	Top 3,000
Dice	190	335	360	375	390
LLR	107	249	311	349	382
TMI	116	257	314	359	382
T-score	211	336	354	372	391

Using these ranking metrics in the Economics corpora most of the 110 reference terms are placed in position 150 (Log-likelihood, True Mutual Information and T-score) on the list of 192 term candidates. In consequence, by validating only the first 150 term candidates, instead of all 192, we obtained 96 terms. Table 3 shows how reference terms from the Economics domain are distributed by these four ranking metrics.

Table 3. Number of reference terms list distribution (Economics)

Method	Top 50	Top 75	Top 100	Top 150
Dice	38	53	67	92
LLR	37	54	70	96
TMI	37	54	70	96
T-score	32	48	71	96

Third, term candidates were reviewed manually. In order to compare results described above, we prepared experimental results filtering term candidates with a new list of reference terms: 500 reference terms (randomly selected from a 1,000 bigram terms list) for the Telecommunications domain and 300 reference terms (randomly selected from a 518 bigram terms list) for the Economics domain.

Following token slot extraction, in the Telecommunications domain, we obtained 3,246 term candidates and 89,182 non-term candidates. Samples of 1,200 term candidates and non-term candidates were corrected manually. After validation of the term candidates sample, we obtained 664 new terms, i.e., terms not in the reference terms lists (55.33% precision) and 259 reference terms (21.58% precision). Indeed, after validation of the non-term candidates sample we obtained 244 terms. Thus, the algorithm achieved precision of 76.92% at recall of 79.09% (F-measure= 77.99%). Figure 5 (a) shows how 923 terms selected manually are distributed by ranking metrics.

In the Economics domain, we obtained 195 term candidates and 2,019 non-term candidates. All of the term candidates were corrected manually and a sample of 500 non-term candidates was corrected manually. After validation of term candidates, we obtained 117 terms. And after validation of the non-term candidates sample we obtained 49 terms. Thus, the algorithm achieved precision of 60% at recall of 70.48% (F-measure= 64.82%). Figure 5 (b) shows how 117 terms selected manually are distributed by ranking metrics. New terms selected manually were used to enrich the reference terms lists and to select more term candidates during the token slot extraction step

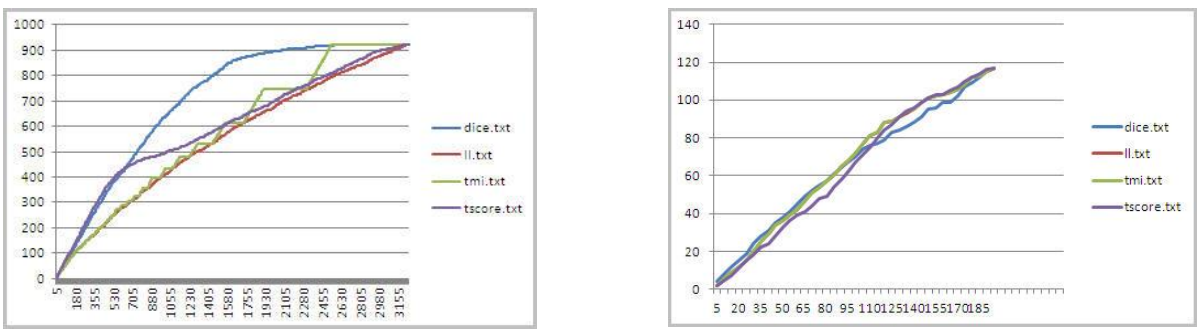


Figure 5. (a) 923 reference terms list distribution (Telec); (b) 117 reference terms list distribution (Economics)

Using these ranking metrics in the Telecommunications corpora most of the 923 terms reviewed manually are placed between positions 2,000 (Dice) and 2,500 (T-score and True Mutual Information) on the list of 3,246 term candidates. In consequence, by validating only the first 2,500 term candidates, instead of all 3,246, we obtained 918 reference terms. Table 4 shows the number of terms distributed by these four ranking metrics.

Table 4. Manual term candidates review: Telecommunications distribution

Method	Top 500	Top 1,500	Top 2,000	Top 2,500	Top 3,000
Dice	362	817	893	918	923
LLR	237	549	671	785	880
TMI	241	570	745	890	923
T-score	379	590	695	801	899

Using these ranking metrics in the Economics corpora most of the 117 reference terms are placed in position 150 (Log-likelihood, True Mutual Information and T-score) on the list of 195 term candidates. In consequence, by validating only the first 150 term candidates, instead of all 195, we obtained 101 terms. Table 5 shows how manual terms selected from Economics domain are distributed by these four ranking metrics.

Table 5. Manual term candidates review: Economics distribution

Method	Top 50	Top 75	Top 100	Top 150
Dice	38	55	70	95
LLR	36	54	72	101
TMI	36	54	72	101
T-score	33	48	67	101

Results obtained show that the token slot extraction step combined with ranking the term candidates improves term candidate validation. In the Telecommunications domain, terms are placed in the best positions using Dice, T-score and True Mutual Information ranking metrics: terms are placed between positions 2,000 (Dice) and 2,500 (T-score and True Mutual Information) on the lists of 3,385 and 3,246 term candidates. In consequence, by validating only the first 2,000 or 2,500 term candidates, instead of all 3,385 and 3,246, we obtained most of the terms (918/923). In the Economics domain, terms are placed in the best positions using Log-likelihood, True Mutual Information and T-score ranking metrics: terms are placed in position 150 on the lists of 192 and 195 term candidates. In consequence, by validating only the first 150 term candidates, instead of all 192 and 195, we obtained most of the terms (96/117).

4. Conclusion

To conclude, we present significant results to improve term candidate validation (time and precision). Using token slot extraction we obtained terminological patterns for term candidates extraction and reduced the list of candidates to be validated manually, removing those candidates that are not terms. More specifically, from a list of 92,428 term candidates in the Telecommunications corpora and 2,214 term candidates in the Economics corpora, the algorithm filtered 3,385 and 192 term candidates respectively, those that have one or more slots that could be filled by terminological tokens from the reference terms. Likewise, the algorithm proposed improves term candidates extraction using those candidates selected as terms to filter term candidates which have not been taken into account in the first token slot extraction step, and so on, recursively.

Furthermore, an evaluation done using reference terms showed which ranking metrics improves results and reduces the number of candidates that need to be evaluated manually. In the Telecommunications domain, Dice, T-score and True Mutual Information obtained the best results. In the Economics domain, Log-likelihood, True Mutual Information and T-score obtained the best ranking metrics.

Moreover, the evaluation done using manual term candidate's review showed a high precision (76.92%) and recall (79.09%) in the Telecommunications domain. As for the Economics domain,

precision (60%) and recall (70.48%) were a little bit lower than the Telecommunications domain, probably due to the corpus size and also the number of reference terms used to filter out term candidates.

Finally, the algorithm proposed allows extracting term candidates from minority language corpora with poor language processing tools. Therefore, a higher number of corpora can be used to extract new terms from specialized domains, which is very important in the work to identify neologisms, expand terminological databases, monitor terminological evolution in specialized domains or manage information from a specific domain.

References

[1] Foo, J. *Computational Terminology: Exploring Bilingual and Monolingual Term Extraction*. Linköping University, 2012.

[2] Sager, J. C., & Nkwenti-Azeh, B. *A Practical Course in Terminology Processing*. John Benjamins Publishing Company, 1990.

[3] Heid, U., & McNaught, J. EUROTRA-7 Study: *Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications*. IMS, University of Stuttgart, 1991.

[4] Frantzi, K. T., & Ananiadou, S. Automatic Term Recognition Using Contextual Cues. *Proceedings of 3rd DELOS Workshop*, 1997.

[5] Vu, T., Aw, A., & Zhang, M. Term Extraction Through Unithood and Termhood Unification. *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, Hyderabad, India, 2008.

[6] Bourigault, D. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. *Proceedings of the 14th Conference on Computational Linguistics*, 1992, 3, pp. 977–81.

[7] Daille, B. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 1996, 1, pp. 49-66.

[8] Bourigault, D., & Jacquemin, C. Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, 1999, pp. 15–22.

[9] Pazienza, M., Pennacchiotti, M., & Zanzotto, F. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. *Knowledge Mining*, 2005, pp. 255-79.

[10] Lafon, P. *Dépouillements et Statistiques en Lexicométrie*. Genève-Paris, Slatkine-Champion, 1984.

[11] Church, K.W., & Hanks, P. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 1990, 16(1), 22–9.

[12] Calzolari, N., & Bindi, R. Acquisition of Lexical Information: From a Large Textual Italian Corpus. *Proceedings of the 13th Conference on Computational Linguistics*, 1990, 3.

[13] Smadja, F.A., & McKeown, K.R. Automatically Extracting and Representing Collocations for Language Generation. *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, 1990.

[14] Ananiadou, S. *Automatic Term Recognition*. Sophia, 2009.

[15] Nenadic, G., Spasic, I., & Ananiadou, S. Mining Term Similarities from Corpora. *Terminology*, 10(1), 55-80.

[16] Daille, B. Variations and Application-Oriented Terminology Engineering. *Terminology*, 2005, 11(1), 181-97.

[17] Weller, M., Blancafort, H., Gojun, A., & Heid, U. Terminology Extraction and Term Variation Patterns: A Study of French and German Data. *Proceedings of the GSCL: German Society for Computational Linguistics and Language Technology*, Hamburg, Germany, 2011.

[18] Nenadic, G., Ananiadou, S., & McNaught, J. Enhancing Automatic Term Recognition through Recognition of Variation. *Proceedings of the 20th International Conference on Computational Linguistics*, 2004, 604.

[19] Gojun, A., Heid, U., Weissbach, B., Loth, C., & Mingers, I. Adapting and Evaluating a Generic Term Extraction Tool. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012.

[20] Oliver, A., Vázquez, M., & Moré, J. Linguoc Lexterm: Una Herramienta de Extracción Automática de Terminología Gratuita. *Translation Journal*. Available from: http://learningtechnologies.uoc.edu/resources/

[21] Moré, J., Vázquez, M., & Villarejo, L. Lexterm, An Open Source Tool for Lexical Extraction. *IV International Seminar on Natural Language Processing, Computational Lexicography and Terminology*. Slovak Academy of Sciences, Bratislava, 2007. Available from: http://learningtechnologies.uoc.edu/resources/

[22] Oliver, A., & Vàzquez, M. A Free Terminology Extraction Suite. *Translating and the computer,* 2007, 29. ASLIB. London.

[23] Moré, J., Rius, L., Vázquez, M., & Villarejo L. L'observatori de Terminologia Talaia: Mètode i Processos. *Tradumàtica: Traducció i Tecnologies de la Informació i la Comunicació,* 2008.

[24] Montes, D., & Vázquez, M. L'observatori de Terminologia Talaia: Un Exemple D'innovació per Mitjà de la Cooperació. *Llengua i ús: revista tècnica de política lingüística*, 2009, 44, pp. 26–35.

[25] Merkel, M., & Foo, J. Terminology Extraction and Term Ranking for Standardizing Term Banks. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA-2007)*, 2007.

[26] Banerjee, S., & Pedersen, T. The Design, Implementation, and Use of the Ngram Statistics Package. *Computational Linguistics and Intelligent Text Processing*, 2003, pp. 370–81.

[27] Wermter, J., & Hahn, U. Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 843-50.

[28] Evert, S. *The Statistics of Word Co-occurrences*. *Word Pairs and Collocations*. PhD thesis, Stuttgart, 2005.

[29] Evert, S., & Krenn, B. Association Measures, [Online] 2004. Available from: http://www.collocations.de/AM/index.html.

[30] Crater. Corpus Resources and Terminology Extraction, [Online] 1995. Available from: http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html.

[31] IULA corpus. Technical Corpus from Institute for Applied Linguistics (IULA), [Online] 2011. Available from: http://bwananet.iula.upf.edu/indexen.htm

[32] *Diccionari de telecomunicacions*. Barcelona: Enciclopèdia Catalana. TERMCAT, Centre de Terminologia, 2007.

[33] Krauthammer, M., & Nenadic, G. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics*, Elsevier, 2004.