# HOW STATISTICAL INFORMATION FROM THE WEB CAN HELP IDENTIFY NAMED ENTITIES

Mathieu Roche
*LIRMM, CNRS, Univ. Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5, France*

Keywords: Text-mining, Natural language processing, Terminology, Named entity.

Abstract: This paper presents a Natural Language Processing (NLP) approach to filter Named Entities (NE) from a list of collocation candidates. The NE are defined as the names of 'People', 'Places', 'Organizations', 'Software', 'Illnesses', and so forth. The proposed method is based on statistical measures associated with Web resources to identify NE. Our method has three stages: (1) Building artificial prepositional collocations from Noun-Noun candidates; (2) Measuring the "relevance" of the resulting prepositional collocations using statistical methods (Web Mining); (3) Selecting prepositional collocations. The evaluation of Noun-Noun collocations from French and English corpora confirmed the relevance of our system.

## 1 INTRODUCTION

In this paper, we focus on the study of French and English phrases called collocations that can be extracted using NLP (Natural Language Processing) methods. (Clas, 1994) uses two properties to define a collocation. First, it is defined as a phrase with a meaning that can be deduced from the words that comprise the phrase. For instance, the French phrase *lumière vive* (*bright light*) is considered to be a collocation because the overall meaning of the phrase can be deduced from the two words *lumière* and *vive*. Based on this definition, the phrase *tirer son chapeau* (*to take off one's hat*), is not a collocation because its meaning cannot be deduced from each word. These forms are called "fixed expressions".

A second property was added by the author (Clas, 1994) to define a collocation. The meaning of the words of the collocation has to be limited. For example, *acheter un chapeau* (*buy a hat*) is not a collocation since the meaning of *acheter* and *chapeau* is not limited. Indeed, a multitude of objects can be purchased. Such phrases are called "free expressions". However it is still very difficult to distinguish between "fixed expressions, "free expressions", and collocations automatically.

The above definitions of collocation can be enriched with two additional characteristics, i.e. semantic and syntactic knowledge (Heid, 1998). The first point is based on a semantic shared by the collocations. For example, *lait tourné* (*sour milk*) and *beurre rance* (*rancid butter*) have a very similar meaning as both describe a degradation phenomenon. The semantic information contained in collocations has been taken into account in many studies (Melcuk et al., 1999; Heid, 1998). The second linguistic characteristic is the syntactic information contained in the collocation (Clas, 1994; Daille, 1996). Generally, collocations have one of the following patterns: Noun-Verb, Noun-Adjective, Noun-Noun, Noun-Preposition-Noun, Verb-Adverb, Adverb-Adjective, etc.

Using automatic processing, it is difficult to identify collocations based on all these linguistic definitions. So we decided to focus on the extraction of "collocation candidates" displaying a specific syntactic pattern. However, the originality of the approach described in this paper is the automatic identification of Named Entities (NE) among these candidates.

NE are conventionally defined as the names of 'People', 'Places', and 'Organizations'. Originally, this definition was used in evaluation challenges such as MUC (Message Understanding Conferences). Today these kinds of challenges (e.g. TREC in English, DEFT in French) cover a wide range of tasks. As indicated in (Daille et al., 2000), the basic classes of NE defined in MUC challenges needed to be enriched. For example, (Paik et al., 1994) defined two new classes: 'Document' (e.g. software, hardware) and 'Science' (e.g. illness, medication). To identify NE, many systems rely on the use of uppercase letters (Farkas et al., 2007). However, this may not be appli-

Roche M..
HOW STATISTICAL INFORMATION FROM THE WEB CAN HELP IDENTIFY NAMED ENTITIES.
DOI: 10.5220/0003473906850689
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WTM-2011), pages 685-689
ISBN: 978-989-8425-51-5

cable to non-capitalized NE (for instance in non-standard documents like emails, blogs, tweets, texts or text fragments that are written using either uppercase or lowercase). Consequently, in the present study, we did not chose to exploit this lexical information to identify NE. Neither do we use machine learning approaches for the recognition of NE cited in (Baluja et al., 2000; Farkas et al., 2007).

In addition to characterize the NE, the criteria *uniqueness referential* (i.e. a proper name refers to a single referential entity), and *stability of the denomination* (i.e. few possible variations) are specified by (Fort et al., 2009). Our approach is based on this last criterion to identify NE among collocation candidates. The extraction of collocation candidates is based on a text mining process (see 'stage 1' – Figure 1). After the acquisition of a corpus, a cleaning process has to be applied. This removes noise in the texts (e.g. HTML tags) that can lead to mistakes in an NLP application. With the normalized texts, a part-of-speech (PoS) tagger can be used. This attributes a grammatical label (e.g. adjective, noun, and so on) to each word of the corpus (Brill, 1994). In this paper, we describe a method to extract collocation candidates from tagged texts. The process is based on the use of patterns to extract nominal terminology such as Noun-Adjective (in French), Adjective-Noun, Noun-Noun, Noun-Preposition-Noun (Bourigault and Jacquemin, 1999; Daille, 1996; Roche and Kodratoff, 2006).

In the following section (section 2), we focus on a statistical method for selecting NE from Noun-Noun collocation candidates (see 'stage 2' – Figure 1). The results of experimental tests of our method conducted on real data (French and English corpora) are presented in section 3. Finally, section 4 presents our future work.

# 2 NAMED ENTITY FILTERING

## 2.1 General Principle

Noun-Noun terminology has often various forms. For example, the French collocation *fichier clients* (*customer file*) can be found with the Noun-Preposition-Noun form: *fichier de clients* (*file of customers*), *fichiers pour clients* (*file for customers*), and so forth. Note that the variations of NE are rare. Based on this remark we will use NLP methods in order to identify NE from a list of Noun-Noun collocation candidates.

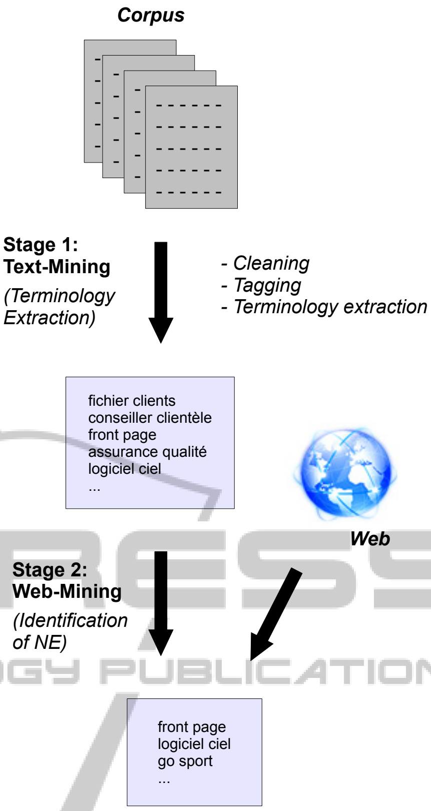

Figure 1: Global process.

Our method has three stages detailled in the following section:

1. Building artificial prepositional collocations from Noun-Noun candidates.
2. Measuring the "relevance" of the resulting prepositional collocations using statistical methods.
3. Selecting prepositional collocations with low scores (i.e. irrelevant built collocation).

## 2.2 Description of the Process

**Step 1 - Building.** Based on Noun-Noun candidates, we built prepositional candidates by adding the most common preposition: "de" (in French[1]) and "of" (in English). When this principle was applied to both French examples *fichier clients* (*customer file*) and *logiciel ciel* (*ciel software*), we obtained the following results:

- fichier clients $_{NN}$ → fichier de clients $_{N-Prep-N}$
- logiciel ciel $_{NN}$ → logiciel de ciel $_{N-Prep-N}$

[1] Note that when the second word of the candidate begins with a vowel, the preposition used is " d' ".

Note that the English variations of Noun-Noun candidates consist in adding a preposition associated with a swapping operation between the nouns (Jacquemin, 1997):

- knowledge discovery $_{NN}$ → discovery of knowledge $_{N-Prep-N}$

**Step 2 - Measuring.** The goal of the second step was to measure the dependence between each word of the collocations we built. This process is based on the use of the Dice measure (Smadja et al., 1996) which has good behavior (Roche and Prince, 2008; Roche and Kodratoff, 2009). The measure is defined by the following formula based on three elements as in (Petrovic et al., 2006):

$$Dice(x,y,z) \quad = \quad \frac{3 \times nb(x,y,z)}{nb(x) + nb(y) + nb(z)} \qquad (1)$$

The principle of this measure is to calculate the number of occurrences of each word $x$ (i.e. $nb(x)$) or collocation $x\ y\ z$ (i.e. $nb(x,y,z)$). In general, the number of occurrences represents the frequency of the words and/or collocations. This frequency is generally calculated regarding a corpus (Daille, 1996). In our case, the Dice measure is applied in a web-mining context (Turney, 2001).

Thus, the frequency *nb* corresponds to the number of web pages containing words or collocations. This number is returned by querying search engines (Google, Yahoo, Exalead, etc.). For example, $nb(fichier)$ is the number of pages returned with the single keyword, and $nb(fichier, de, clients)$ is the number of pages returned with the query *"fichier de clients"* (by using quotes to search an exact phrase). An example of values obtained with Dice measure is given below:

$$Dice(fichier, de, clients) = \frac{3 \times 999{,}000}{37{,}200{,}000 + 6{,}350{,}000{,}000 + 208{,}000{,}000} = 0.000454$$

$$Dice(logiciel, de, ciel) = \frac{3 \times 89{,}800}{35{,}000{,}000 + 6{,}350{,}000{,}000 + 35{,}400{,}000} = 0.0000419$$

This result shows that the lowest score in major proportions (factor ten) is given by *logiciel ciel.* Thus, our measure predicts that this Noun-Noun candidate is a NE. This is very relevant because this one is a management software adapted to one class of NE (i.e. 'Document' class) (Daille et al., 2000; Paik et al., 1994). Web gives an indication of popularity of the words/collocations. Moreover with "external" knowledge (i.e. Web) we are less sensitive to the size of the used data (i.e. corpus).

Note that this measure has two differencies with the web-mining approach described in (Turney, 2001): (1) We use the Dice measure because it provides best results than Mutual Information (Roche and Kodratoff, 2009); (2) The numerator of this measure is based on the exact search of phrases (by using quotes) and not on the presence (i.e. AND operator) or the proximity (i.e. NEAR operator) of the nouns.

**Step 3 - Selecting.** Constructed candidates with low scores are elements with few possible variations. In our approach, these candidates are considered to be NE. We introduce a parameter $S$ which represents a selection threshold. For example, with a threshold $S = 10$, ten candidates with the lowest scores will be selected as possible NE. The results with different values of $S$ are discussed in the following section.

# 3 EXPERIMENTS

## 3.1 Experimental Protocol

In our experiments, the first used corpus (called CV) is composed of $1,144$ Curriculum Vitae provided by the VediorBis company ($120,000$ words). The second specialized French corpus (called HR) is composed of a set of texts from the Human Resources field. The texts correspond to summaries of psychological tests of 378 persons ($600,000$ words). First we select collocation candidates which are present a minimum number of times in the corpus. This pruning task enables to exclude candidates which are often unrepresentative of the field. The Table 1 shows the number of candidates obtained before and after pruning at 3 such as (Thanopoulos et al., 2002; Roche and Kodratoff, 2006).

According to the same field and language, the results may differ. For example, with the CV corpus after pruning, the number of Noun-Noun candidates is much larger than the Human Resources corpus. Yet the size of Human Resources corpus is five times larger than the CV corpus. This is because CV are written in a condensed way, using very specific nominal terminology. We select this kind of collocation in the study presented below.

The goal of our experiments is to evaluate if the collocation candidates selected with our web-mining approach represent really relevant NE. In this context, we relied on the most frequent Noun-Noun candidates of the CV corpus. In our experiments, 70 candidates have been evaluated manually (18 NE have been identified). Note that this protocol needs an automatic execution of 210 queries[2] with the search engine Exalead

[2] We apply 70 queries for the numerators, and 140

Table 1: Number of collocation candidates obtained before and after pruning.

| | Before pruning | | After pruning | |
|---|---|---|---|---|
| | *HR* | *CV* | *HR* | *CV* |
| *Noun-Noun* | 98 | 1,781 | 11 | 162 |
| *Noun-Preposition-Noun* | 4,703 | 3,634 | 1,268 | 307 |
| *Adjective-Noun* | 1,260 | 1,291 | 478 | 103 |
| *Noun-Adjective* | 5,768 | 3,455 | 1,628 | 448 |

Table 2: Precision, Recall, and F-measure with different values of $S$ – French corpus.

| | | French corpus | | |
|---|---|---|---|---|
| | *S* | *10* | *40* | *70* |
| Web-Mining ranking | Precision | **0.60** | 0.35 | 0.26 |
| | Recall | 0.33 | 0.78 | **1** |
| | *F-measure* | *0.43* | ***0.48*** | *0.41* |
| Random ranking | *F-measure* | *0.18* | *0.35* | *0.41* |

Table 3: Precision, Recall, and F-measure with different values of $S$ – English corpus.

| | | English corpus | | | | | |
|---|---|---|---|---|---|---|---|
| | *S* | *10* | *40* | *70* | *140* | *210* | *305* |
| Web-Mining ranking | Precision | **1** | 0.83 | 0.73 | 0.58 | 0.46 | 0.34 |
| | Recall | 0.10 | 0.31 | 0.49 | 0.77 | 0.92 | **1** |
| | *F-measure* | *0.17* | *0.46* | *0.58* | ***0.66*** | *0.62* | *0.51* |
| Random ranking | *F-measure* | *0.06* | *0.19* | *0.28* | *0.39* | *0.46* | *0.51* |

(http://www.exalead.com/), this search engine giving good results, in particular for French data.

To evaluate the generalisation of our approach, we also propose to work with an English corpus composed of Noun-Noun collocation candidates. This list has 105 NE (examples of topics: politicy, military, religious, etc) and 200 terms (topic: data-mining). We have ranked all the candidates applying our web-mining approach by performing 915 queries.

The benchmarks used in these experiments are very specific because they have only lowercase words. This situation is distinct from classic benchmarks of the state-of-the-art.

### 3.2 Results

In order to evaluate our text-mining system Precision (formula (2)) and Recall (formula (3)) are calculated. A precision at 100% means that all NE returned with our system are relevant, and a recall at 100% means that all relevant NE are returned. In addition, the F-measure (formula (4)) combines the precision and recall.

$$Precision = \frac{\text{Nb of returned relevant NE}}{\text{Nb of returned NE}} \qquad (2)$$

$$Recall = \frac{\text{Nb of returned relevant NE}}{\text{Nb of relevant NE}} \qquad (3)$$

$$F-measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

We measure the quality of our system with different thresholds $S$ (see Tables 2 and 3). The results show that the best value of F-measure is obtained when we take into account the first half of candidates from French and English corpora.

We can consider that our method has a good behaviour because our ranking function returns a lot of NE at the top the list. For instance, with the English corpus, the value of the precision is 83% when 40 candidates are selected with our web-mining approach (a random ranking returns a precision at 34%).

Finally, note that the F-measure with our approach is better in a large proportion than a random distribution for all values of $S$ (see Tables 2 and 3).

queries for both nouns of the denominators (the query for the prepositions have been applied only once for all calculations).

## 4 CONCLUSIONS AND FUTURE WORK

This paper presents a text mining method for extracting collocation candidates and identifying Named Entities from a list of candidates. The filtering method uses only a statistical approach based on the Dice measure and exploitation of the results of search engines. The NE is "stable", which is why we built variations of the candidates and checked their popularity using search engines. If the candidates we built turned out to be irrelevant (i.e. low value of the statistical measure), they were considered as NE.

In the next step of our work, we plan to enrich the rules concerning variations, which is a precondition to take into account the vast majority of possible linguistic variations. Finally, we plan to combine this approach, which is only based on statistical knowledge, with lexical information, particularly the use of uppercase letters when possible.

## ACKNOWLEDGEMENTS

I thank Daphne Goodfellow who improved the readability of this paper.

## REFERENCES

Baluja, S., Mittal, V. O., and Sukthankar, R. (2000). Applying machine learning for high-performance named-entity extraction. *Comput. Intelligence*, 16(4):586–596.

Bourigault, D. and Jacquemin, C. (1999). Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 15–22.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *Proceedings of AAAI (Conference on Artificial Intelligence)*, volume 1, pages 722–727.

Clas, A. (1994). Collocations et langues de spécialité. *Meta*, 39(4):576–580.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press*, pages 49–66.

Daille, B., Fourour, N., and Morin, E. (2000). Catégorisation des noms propres : une étude en corpus. *Cahiers de Grammaire*, 25:115–129.

Farkas, R., Szarvasand, G., and Ormandi, R. (2007). Improving a state-of-the-art named entity recognition system using the world wide web. In *Proceedings of Industrial Conference on Data Mining*, pages 163–172.

Fort, K., Ehrmann, M., and Nazarenko, A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus. In *Proceedings of TALN (Traitement Automatique du Langage Naturel)*.

Heid, U. (1998). Towards a corpus-based dictionary of german noun-verb collocations. In *Proceedings of the Euralex International Congress*, pages 301–312.

Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. In *Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes*.

Melcuk, I., Arbatchewsky-Jumarie, N., Elnitsky, L., and Lessard, A. (1984, 1988, 1992, 1999). Dictionnaire explicatif et combinatoire du français contemporain. *Presses de l'Université de Montréal*, 1,2,3,4.

Paik, W., Liddy, E., Yu, E., and McKenna, M. (1994). Categorizing and standardizing proper nouns for efficient information retrieval. In *Corpus Processing for Lexical Acquisition, MIT Press, chap. 4*.

Petrovic, S., Snajder, J., Dalbelo-Basic, B., and Kolar, M. (2006). Comparison of collocation extraction measures for document indexing. In *Proceedings of ITI (Information technology interfaces conference)*, pages 451–456.

Roche, M. and Kodratoff, Y. (2006). Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *Proceedings of onToContent Workshop - OTM'06, Springer Verlag, LNCS*, pages 1107–1116.

Roche, M. and Kodratoff, Y. (2009). Text and web mining approaches in order to build specialized ontologies. *Journal of Digital Information (JoDI)*, 10(4).

Roche, M. and Prince, V. (2008). Managing the Acronym/Expansion Identification Process for Text-Mining Applications. *International Journal of Software and Informatics*, 2(2):163–179.

Smadja, F., McKeown, K., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons : A statistical approach. *Comp. Linguistics*, 22(1):1–38.

Thanopoulos, A., Fakotakis, N., and Kokkianakis, G. (2002). Comparative evaluation of collocation extraction metrics. In *Proceedings of LREC (International Conference on Language Resources and Evaluation)*, pages 620–625.

Turney, P. (2001). Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In *Proceedings of ECML/PKDD (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases)*, pages 491–502.