Generating a Resource for Products and Brandnames Recognition. Application to the Cosmetic Domain.

Cédric Lopez*, Frédérique Segond*, Olivier Hondermarck,**
Paolo Curtoni*, Luca Dini*****

*Viseo Research Center, Grenoble (France)
**Beaute-test.com
*** Holmes Semantic Solutions, Grenoble (France)
clopez@objetdirect.com, fsegond@objetdirect.com, webmaster@beaute-test.com, curtoni@ho2s.com, dini@ho2s.com

Abstract

Named Entity Recognition task needs high-quality and large-scale resources. In this paper, we present RENCO, a based-rules system focused on the recognition of entities in the Cosmetic domain (brandnames, product names, …). RENCO has two main objectives: 1) Generating resources for named entity recognition; 2) Mining new named entities relying on the previous generated resources.
In order to build lexical resources for the cosmetic domain, we propose a system based on local lexico-syntactic rules complemented by a learning module. As the outcome of the system, we generate both a simple lexicon and a structured lexicon. Results of the evaluation show that even if RENCO outperforms a classic Conditional Random Fields algorithm, both systems should combine their respective strengths.

Keywords: Named Entity Recognition, Name of Products, Brandnames

1. Introduction

Named Entity Recognition (NER) is an essential task in the field of Information Extraction (IE) which includes challenges such as disambiguation and fine grained categorization. In recent evaluation campaigns such as Ester 2 or Automatic Content Extraction (ACE), respectively, 6 and 7 main classes were defined, representing a total of fifty elements (subclasses). Nowadays, it is rather usual to use up to hundreds of classes to describe named entities. In their work (Sekine and al., 2002) proposed approximately 150 categories of named entities, that they enriched later to come up to a total of 200 in (Sekine and Nobata, 2004).

For instance, Politicians is a sub-category of Persons (Seungwoo and Gary, 2005), city names (Fleischman, 2001), anthroponyms (Fourour, 2001) and the place-names (Fourour on 2002) are sub-categories of Localisation. In the same way, depending on the domain specialisation and needs for the IE task, new categories appear, such as names of genes or proteins in the field of the microbiology (Kim and al., 2004; Hirschman and al., 2005), or name of products

While NER is widely studied, there are few publications dedicated to the recognition of product and brand names (Zhao, 2008). However, with the development of applications such as opinion analysis or personalised marketing, being able to detect brands and product names becomes crucial.

Even if names of products and brandnames are well integrated in current typologies, few resources are available. When it comes to recognising product or brand names, compared to other types of named entities, the main issue is the very high level of ambiguity. As pointed out in Díaz (2001), in his study about perfume onomastic, ambiguity is even stronger in the cosmetic domain where any words could be considered as a name of perfume. For instance, a perfume name in the cosmetic domain can be: a number (*N°5)*, a date (*1881*), an address (*24 Faubourg*)*,* a sentence (*La vie est belle*), a pronoun (*Elle*), and so forth. Such a large spectrum for product names that can turn any common noun phrase into a product or brand name makes the task of building specific lexical resources very difficult, and as a consequence, to our knowledge, there are no available lexical resources relative to the cosmetic domain.

In this paper we describe our system, called RENCO (REcognition of Named entity in COsmetic), which has two in linked objectives: to recognize named entities, and to generate lexical resources (without preliminary manual annotation) related to the cosmetic domain. Hence, RENCO is at the frontier of Named Entity Recognition and Taxonomy Extraction (see Medelyan et al., 2013 for an overview of automatic construction of resources). Both tasks overlap, since large coverage specialized lexicons constitute the foundation of any good NER system (McDonald, 1994; Wakao et al., 1996). In this way, our work is inspired by (Hearst, 1992) as it is based on the automatic construction of lexical resources. Indeed, our system is based on lexico-syntactic rules. It is easily adaptable for the generation of more general resources, at least within the framework of products and brand names (not limited to cosmetics).

In what follows we first present the adopted typology for brand and product names (section 2.1), we then introduce the lexico-syntactic rules (section 2.2) that are used in our global system (section 2.3). We conclude with a comparison between RENCO and a classical statistical method (Conditional Random Fields) in section 3 and we suggest future directions for our research.

2. Recognition of Brand and Product Names

In order to build lexical resources for the cosmetic domain, we propose a system based on local grammatical rules complemented by a learning module. The outcome of the system is both:

1. A simple lexicon (not structured): COSM
2. A lexicon structured in agreement with our typology (see below): COSM-XML

In this study, we use a corpus of 1,000 French articles gathered from four magazines specialized in the cosmetic field (Beauté Infos, Féminin Pratique, Cosmétique Hebdo, Cosmétique Mag).

2.1 Typology

We started from the following classes of named entities to model lexical items of the cosmetic domain: Names of products, range of products, brand names, divisions, and groups (see examples on Fig. 1).
Co-occurrences study in our corpus shows direct relations between some classes presented above. Based on the meronymic relation (for example using prepositions as “de”, “chez”, or brackets), we propose 5 levels hierarchy adapted to named entity recognition (cf. Fig. 1) easily adaptable to other products out of the cosmetic domain (for instance in the automotive sector or food domain). In figure 1, thick arrows indicate frequent co-occurrences in our corpus, showing that brand names are playing a “pivot” role between Group, Division, Range, and Product.

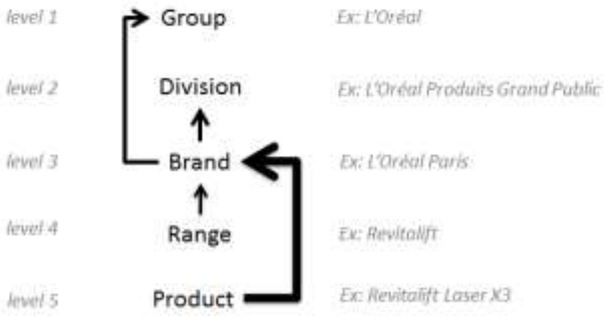

Figure 1: Hierarchy of cosmetic entities adapted to named entity recognition

Based on the defined hierarchy, we define in the next section the rules we set up.

2.2 Evidence Rules for Brand/Product Recognition

Starting from the hierarchy presented in the previous section (cf. Fig. 1), we define different types of evidence rules to recognise named entities in the cosmetic domain. These rules are applied to noun phrases only (we favour local context in order to avoid noise when generating the resource).

- **Lexico-syntactic evidence rules**

- D*efining rules* use the principle of external evidence (Mc Donald, 1994). Such rules are based on hyponymic and hyperonymic relations (for example "the flavor such as Angel"). In order to extract such relations, we rely on lexico-syntactic patterns. Inspired by (Hearst, 1992) we adapt generic patterns as “such NP as {NP,}*” that becomes “*brandnames* such as *NE*” where *NE* is a named entity that belongs to the brand name category.

For instance, in the sentences *Les marques telles que Lancôme et Guerlain* (*Brands such as Lancôme and Guerlain*) or *Prodigieuse Nuit est une crème de nuit hydratante* (*Prodigieuse Nuit is a hydratant night cream*) descriptors like *telles que* (*such as*) and *est une (is a*) enable to deduce that “Lancôme” and “Guerlain” are brandnames and “Prodigieuse Nuit” is a product.

- I*nternal contextual rules* use the principle of internal evidence (Mc Donald, 1994) where internal evidence is a term included in an entity, that enable the annotation with a strong reliability. For example "Clarins Fragrance Group" where "Group" indicates clearly the type of this named entity.

Furthermore, we consider our hierarchy to define new internal contextual rules. In particular, according to the defined hierarchy, we assume that if an entity (noted EN) contains another entity of level *n*, then EN belongs to a lower level (< *n*). For instance, “L’Oréal Luxe” contains “L’Oréal” which is a group (level 1), then we can deduce that “L’Oréal Luxe” (which is a division) might be typed with a lower level (< *n*).

- **Syntactic evidence rules**

- The *rules of coordination* are based on syntactic analysis of the noun phrase and rely on the linguistic fact that in coordination, coordinates are of the same nature. We focus on the most reliable conjunctions “et” (*and*) and “ou” (*or*). For example, knowing that L’Oréal Paris is a brand name, the coordination *and* enables to deduce that Cadum is also a brandname when it appears in: "L’Oréal Paris et Cadum".

- The *hierarchical rules* are based on semantics of prepositions and enable to structure the extracted data (for example "Perfect Mousse est un produit de Schwarzkopf" / "Perfect Mousse is a product by Schwarzkopf "). Prepositions such as “de” (*of, by*) or “chez” (*to*) establish a hierarchical relation (belonging, possession), between two entities (*e.g.* Nastase and Strube, 2008). For instance, knowing that Item Dermatologie is a brand name, the preposition “de” in *j’ai testé Alphadoux de Item Dermatologie* (*I tested Alphadoux of Item Dermatologie*) enables to deduce that Alphadoux is a product.

Defining rules, *internal contextual rules*, and *rules of coordination* enable generating of the COSM lexicons.

Hierarchical rules are applied to generate the structured lexical resource COS-XML, based on the hierarchy presented in Figure 1. Some examples of rules are presented below:

a- Product + Prep (de|d'|chez) + EN => EN= Brand
b- Brand + Prep (de|d'|chez) + EN => EN= Group
c- Range + Prep (de|d'|chez) + EN => EN= Brand
d- Division + Prep (de|d'|chez) + EN => EN= Group
e- EN + Prep (de|d'|chez) + Brand => EN= Product
f- EN + Prep (de|d'|chez) + Range => EN= Brand

In this way, if we have "Perfect Mousse de Schwarzkopf", knowing that "Perfect Mousse" is a product we can deduce that "Schwarzkopf" is a brand name (thanks to the rule *a)*.

The activation of lexico-syntactic evidence rules and hierarchical rules only depends on the presence of an evidence in the context of the entity. These rules do not depend on each other. Conversely, *hierarchical rules* are both inter dependent (the activation of a rule can activate another one) and intra dependent (the activation of a rule depends on the presence of an entity already recognized in the immediate context). Thus, our global process might take into account these specificities by determining the priority of the rules.

For that purpose, we made a first qualitative evaluation based on 10% of our corpus, *i.e.* 203 named entities manually annotated. We compute independently the precision and recall for each module of rules.
First results indicate that our rules enable the annotation of 54% of named entities.
More precisely, *defining rules* obtain a precision of 1 indicating that all annotation made with this module are relevant (in our corpus). The recall is about 0.13. *Hierarchical rules* (after projection of a dictionary containing 3,918 brand names) have a precision of 0,96 and a recall of 0,22. Coordinating rules have a precision of 0,80 and a recall of 0,08. Finally internal contextual rules obtain a precision of 0,46.

As we want to favor the precision rather than the recall, we set the following order of rules application:

1- Defining rules
2- Hierarchical Rules
3- Rules of coordination
4- Internal contextual rules

Next section introduces the global process.

2.3 Global Process

In this section, we describe the global process we use to extract named entities related to the cosmetic domain (cf. Fig. 2). We start with two lexicons:

- LexM: A lexicon containing 3,937 brand names of cosmetic, manually extracted from beaute-test.com
- LexG: A lexicon of 21 groups of cosmetic manually extracted from Wikipedia.

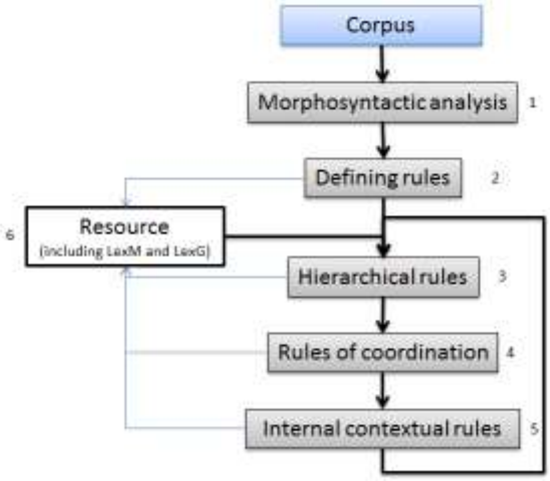

Figure 2: Global Process

The first step consists in analysing the corpus with a morphosyntactic parser provided by Holmes Semantic Solutions[1] in order to detect noun phrases as well as relation among their parts. We then apply defining rules. Steps 3, 4 and 5 consist in applying the rules according to the following order: Hierarchical, coordination and internal evidence rules. The choice of this order is based on the results of the previous qualitative evaluation (*cf.* section 2.2).
Each rule might enrich the resource with new learned entities. All entities (including learned entities) are used in our recognition process. When no entities are detected with previous rules, a projection of LexM and LexG enables to annotate new non-ambiguate brandnames (for example "L'Oréal" is an ambiguous case since it is both a brand name and a group in our lexicons) and might trigger new rules starting again the process from step 3 to 6. The annotated brands can be considered as entities "pivots" in the process (see Fig. 1).

We generate two resources from our corpus:

- COS: A lexicon containing names of products, brand names, ranges of products, divisions and groups of cosmetics, representing 8,7210 named entities (cf. Table 1). From a lexicon containing 3,958 entities we generated 4,773. In other words, COS significantly increased its volume by 120%.

- COS-XML (XML format): A lexicon containing hierarchical links between the NE (see Figure 3), according to our hierarchy presented in section 2.1. For example, *<product name="Kokorico" sup="Jean Paul Gaultier">* indicates that the product Kokorico is proposed by the brand Jean-Paul Gaultier. COS-XML contains 2,081 relations.

[1] http://www.ho2s.com/

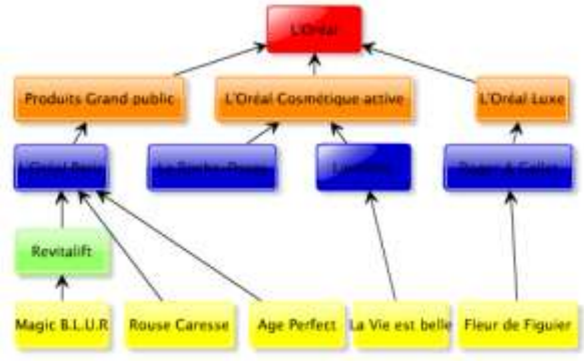

Figure 3: Extract of the COS-XML resource

Types	**Number of entities initially in the resource**	**Number of generated entities**	**Total**
Products	0	2,022	2,022
Brands	3,937	583	4,520
Ranges	0	426	426
Divisions	0	953	953
Group	21	769	800
Total	3,958	4,773	8,721

Table 1: Description of COS

3. Evaluation

This section has two objectives: (1) to evaluate our RENCO system comparing with a baseline, (2) to suggest possible improvement of the system.
First, we describe the evaluation protocol we followed. Then, we compare our symbolic system with a well-known statistic one (CRF). An in-deep evaluation is realised in order to understand the behaviour of our rules. Finally, we discuss the results and propose enhancements.

3.1 Protocol Description

We use a corpus consisting of 1,300 sentences extracted from four French magazines specialized in the cosmetic field (Beauté Infos, Féminin Pratique, Cosmétique Hebdo, Cosmétique Mag). These sentences have been extracted randomly from the magazines, with the only condition that they must contain at least 1 named entity. An expert has manually annotated the corpus in XML format. Then, two sub-corpus have been created:
- “300.xml” which consists of 300 sentences
- “1000.xml” which consists of the 1000 remaining sentences.

A first evaluation consists in comparing our result with a baseline. For this aim, we performed a CRF algorithm with a training having as input the following features:
- Surface word form
- Lemma
- Word shape that is the typographic word composition in terms of uppercase/lowercase characters, numeric digits, symbols...
We created a statistical model trained on the 1000.xml corpus.

We first present the performance of the statistical model applied on the 300.xml "unseen" corpus and we compare with the results of the symbolic system (applied in the same corpus) (see section 3.2). Furthermore, we present a first overview of an in-deep evaluation aiming at enhancing the based-rules system (see section 3.3).

We use precision (P), recall (R), which are classic methods of evaluation in text mining, computed on a representative sample manually annotated.
Precision corresponds to the ratio of the number of entity correctly annotated by the system and the total number of entities annotated by the system. Recall corresponds to the ratio of the number of entity correctly annotated by the system and the number of entity the system should have annotated. We also compute the traditional F-score which is the harmonic mean of precision and recall.

3.2 RENCO *vs.* CRF

In this section we compare our system with the results obtained with a Conditional Random Field (CRF) algorithm. Here, the evaluation is based on single tokens. Each token that is part of a named entity is "classified" with the corresponding named entity type, while the other tokens (not belonging to any NE) are tagged with the special class label "O". Token pairs from the predicted and the manually annotated corpus are aligned and compared. The confusion matrix (see Tables 2 and 3) provides an overview of the predicted/actual classification of corpus tokens. This method implicitly also takes into account the partial recognition of NE. Results in terms of Precision/Recall are presented in Table 4.

	O	Group	Div.	Brand	Range	Prod.
O	**7973**	3	4	13	5	39
Group	7	**152**	4	10	1	0
Div.	26	8	**158**	9	0	4
Brand	58	20	1	**505**	4	9
Range	**55**	0	0	4	17	17
Prod.	135	1	2	6	10	**238**

Table 2: Result of the evaluation on our based-rules system

Globally, both systems obtain satisfactory results. Regarding RENCO, an important gap exists for the range and group names recognition in terms of precision (resp. 0,78 and 0,83) whereas the statistical system obtains good results (0,96 and 0,92). On the contrary, RENCO is better than the CRF algorithm for brand names recognition (0,92 vs. 0,86).
In terms of Recall, RENCO obtains good results. We find a large gap between the two systems, in particular for group, product, and range names recognition (up to 0,45).

Even if our system obtains the best F-scores for each category of named entity, both systems have their strengths and weaknesses. Hence, we plan to construct an hybrid approach (RENCO + CRF) in order to benefit from strengths of both systems.

	O	Group	Div.	Brand	Range	Prod.
O	**8006**	0	0	13	1	17
Group	56	**81**	0	37	0	0
Div.	43	0	**153**	9	0	0
Brand	86	7	4	**490**	0	10
Range	**87**	0	0	14	23	25
Prod.	**217**	0	0	9	0	166

Table 3: Result of the statistical approach

	Precision		Recall		F-score	
	RENCO	CRF	RENCO	CRF	RENCO	CRF
O	**0.97**	0.94	0.99	**1.00**	**0.98**	0.97
Brand	**0.92**	0.86	**0.85**	0.82	**0.88**	0.84
Div.	0.93	**0.97**	**0.77**	0.75	0.84	**0.85**
Group.	0.83	**0.92**	**0.87**	0.47	**0.85**	0.62
Prod.	**0.78**	0.76	**0.61**	0.42	**0.68**	0.54
Range	0.78	**0.96**	**0.49**	0.15	**0.60**	0.27

Table 4: Comparing Precision, Recall, and F-score between RENCO and Conditional Random Fields method.

3.2 In-depth Evaluation

This evaluation consists in analysing the results of RENCO applied on the "300.xml" corpus. As our rules are inter and intra dependent, we have had to apply the dictionary of brandnames in order to bootstrap the system.

Among the 1,099 entities present in the corpus, 787 are from the cosmetic domain. The system has annotated 662 entities for which 582 are correctly annotated (80 was badly annotated). Consequently, the global system reaches a precision of 0,88 and a recall of 0,74.

A recognized Named Entity is considered correct when there is a full match of the NE string (possibly multiword) and NE-type between predicted/actual annotations.

Turning to a more detailed analysis, Table 5 shows the evaluation for the distinct rules modules. The projection of the dictionary enables a relevant annotation for 97% of entities. Results show that our rules obtain a total recall of 0,21.

	DICT	DEF	HIE	COORD	INT
Total Annotations	454	59	39	34	59
Correct Annotations	442	56	31	27	52
Precision	0,97	0,95	0,80	0,80	0,88
Recall	0,56	0,07	0,04	0,03	0,07

Table 5: Results of the evaluation on our based-rules system.

Obviously, the evaluation of the system is equivalent to the evaluation of the resource. More precisely, the evaluation shows that the COS resource (generating with DEF, COORD, and INT) reaches a precision of 0,88. The COS-XML resource reaches a precision of 0,80.

3.3 Enhancing the system

We have identified two main errors caused by the system:

- Errors coming from *defining rules* are mainly due to the plurality of the context. For example, if we have the context "le parfum NE" then we can deduce that NE is a product. However, in the case of "les parfums NE", it could be a brand, *e.g.* "les parfums Dior". Furthermore, it could be a product in the case of coordination: "les parfums L'égoïste et La Vie est Belle sont portés par […]". An easy solution consists in proposing fine grained rules:

a- If the context is in the singular form and only one entity is attached to this context, then the entity is a product

b- If the entity is in the plural form and various entities are attached to the context, then the entity is a brand name.

c- If the context is in the plural form and that various entities are attached, then those entities could be products or brand names. For instance, "Les parfums Dior (resp. Angel) et Chanel (*resp.* Gentlemen Only)" where "parfums" is in a plural form, and Dior and Chanel are brand names (*resp.* Angel and Gentlemen Only are products). This case is undecidable without more contexts.

- Errors coming from *hierarchical rules* are mainly due to the following rule: BRAND (EN) => EN=GROUP. In fact, EN can be a division or a Group. Results of our evaluation shown that in 52% of the case, EN was a group and in 48% it was a division.

4. Conclusion

We briefly presented our based-rules system (called RENCO) for generating lexical resources (COS and COS-XML) for a Named Entity Recognition task in the domain of cosmetics. Even if our approach is applied on French texts, the generated resources are language-independent (named entities do not belong to a particular language). From contextual rules based on a syntactic analysis, we propose two resources in the cosmetic domain.

Evaluation shows that a promising work perspective can be combining RENCO with a CRF model in order to improve the precision of at least three categories of named entities:

Group, division, and range names.

In the same time, we plan to set up new rules to enrich our lexicons. In particular, we plan to study the semantic features of verbs used in such specific domains. Furthermore, we plan to apply our system in other domains such as clothes and automobiles.

Even if textual features on cosmetic names are very different from the traditional named entities (person, location, *etc.*), we plan to train a statistical model using both textual features and the lexico-syntactic rules presented in this paper, and compare the results against a pure rule-based system.

5. Acknowledgements

We thank Mathieu Roche (TETIS and LIRMM, Montpellier (France)) for providing the corpora described in this paper.

6. References

Bick, E. (2004). A named entity recognizer for Danish. In Proc. of *4th International Conf. on Language Resources and Evaluation*, pp. 305-308.

Díaz, M. L. (2001). L'onomastique des parfums. In *Presencia y renovación de la lingüistica francesa* (pp. 215-224). Ediciones Universidad de Salamanca.

Fleischman Michael (2001) Automated subcategorization of named entities. In Proc. of the ACL 2001 Student Research Workshop, pages 25–30.

Fourour, N. (2001). Identification et catégorisation automatiques des anthroponymes du français. Actes, *TALN-Récital*, 1, 441-450.

Fourour, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In *TALN*, Vol. 1, pp. 265-274.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2,* pp. 539-545

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC bioinformatics, 6, S1.

McDonald, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition*, 21-39.

Medelyan, O., Witten, I. H., Divoli, A. and Broekstra, J. (2013), Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. WIREs Data Mining Knowl Discov, 3: 257–279. doi: 10.1002/widm.1097

Nastase, V., & Strube, M. (2008). Decoding Wikipedia Categories for Knowledge Acquisition. In AAAI (pp. 1219-1224).

Pierre, J. M. (2002). Mining knowledge from text collections using automatically generated metadata. In *Practical Aspects of Knowledge Management* (pp. 537-548). Springer Berlin Heidelberg.

Sekine, S., Sudo, K., & Nobata, C. (2002). Extended named entity hierarchy. In *LREC* (Vol. 2).

Sekine, S., & Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In *LREC* pp. 1977-1980.

Wakao, T., Gaizauskas, R., & Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of the 16th conference on Computational linguistics-Volume 1,* pp. 418-423

Zhao, J., & Liu, F. (2008). Product named entity recognition in Chinese text. *Language Resources and Evaluation*, *42*(2), 197-217.