Comparing Named-Entity Recognizers in a Targeted Domain: Handcrafted Rules vs. Machine Learning

Ioannis Partalas, Cédric Lopez and Frédérique Segond
Viseo R.&D., Grenoble, France
firstname.lastname@viseo.com

RÉSUMÉ

La reconnaissance d'entités nommées consiste à classer des objets textuels dans des catégories pré-définies telles que les personnes, les lieux et les organisations. Alors que cette tâche suscite de nombreuses études depuis 20 ans, l'application dans des domaines spécialisés reste un défi important. Nous avons développé un système à base de règles et deux systèmes d'apprentissage pour résoudre la même tâche : la reconnaissance de noms de produits, de marques, *etc.*, dans le domaine de la Cosmétique, pour le français. Les systèmes développés peuvent ainsi être comparés dans des conditions idéales. Dans ce papier, nous présentons nos systèmes et nous les comparons.

ABSTRACT

Comparing Named-Entity Recognizers in a Targeted Domain : Handcrafted Rules vs. Machine Learning

Named-Entity Recognition concerns the classification of textual objects in a predefined set of categories such as persons, organizations, and localizations. While Named-Entity Recognition is well studied since 20 years, the application to specialized domains still poses challenges for current systems. We developed a rule-based system and two machine learning approaches to tackle the same task : recognition of product names, brand names, *etc.*, in the domain of Cosmetics, for French. Our systems can thus be compared under ideal conditions. In this paper, we introduce both systems and we compare them.

MOTS-CLÉS : NER, e-Commerce, rule-based system, machine learning.

KEYWORDS: NER, e-Commerce, système à base de règles, système d'apprentissage.

1 Introduction

The goal of Named-Entity Recognition (NER) is to classify textual objects in a predefined set of categories like for example persons, organizations and localizations. The evaluation of NER tools in campaigns such as MUC, CONLL03, ACE 2005 and 2008, suggests that NER is a solved problem. However, as highlighted by (Marrero *et al.*, 2013), the set of named entity types is very limited in such campaigns (and annotated corpora are also limited) and it is important to highlight the difficulties that NER systems face when dealing with entities in a fine-grained way (Sekine & Nobata, 2004).

In particular, with the development of e-Commerce and personalized marketing, new brands, product names, and other relative entities, appear daily ; being able to detect such entities becomes crucial in several applications. For example, on-line market-places (Amazon, eBay) need to extract features about products provided in free text from different sellers. While NER is widely studied, there are few works dedicated to the recognition of name of products and brands (Ritter *et al.*, 2011; Zhao & Liu, 2008). In this work, we decided to focus on the recognition of fined-grained entities, namely products, name of ranges, brands, divisions, and groups in the domain of Cosmetics. A major difficulty comes from the fact that in the cosmetic's domain there is a high level of ambiguity. As indicated by Díaz (2014), in his study about perfume onomastic, ambiguity is even stronger in the cosmetic domain where any word could be considered as a name of perfume. For instance, perfumes can take form of a number (*N^{o}5*), a date (*1881*), an address (*24 Faubourg*), a sentence (*La vie est belle*), or a pronoun (*Elle*). Such a large spectrum for product names that can turn any common noun phrase into a Cosmetics entity makes the task of building specific lexical resources very difficult.

Early, works on NER showed that large coverage specialized lexicons constitute the foundation of any good NER system (McDonald, 1996) (Wakao *et al.*, 1996). The lack of resources related to Cosmetics makes hard to develop systems that can achieve high performances. Hence, recent works focused on symbolic approaches based on hand-crafted linguistic rules which alleviate partially the scarcity problem of annotated data (Bick, 2004) (Lopez *et al.*, 2014). On the other hand, in the framework of machine learning, NER is casted as a sequence labeling problem and is solved by structured methods. For example, Conditional Random Fields (CRFs), are well-known to capture high-level semantics and also can cope with grammatically degraded text, like in short texts, but are subject to the existence of a significant volume of annotated data (Lafferty *et al.*, 2001). Then, which approach should we adopt when dealing with named entity recognition in a specialized domain ?

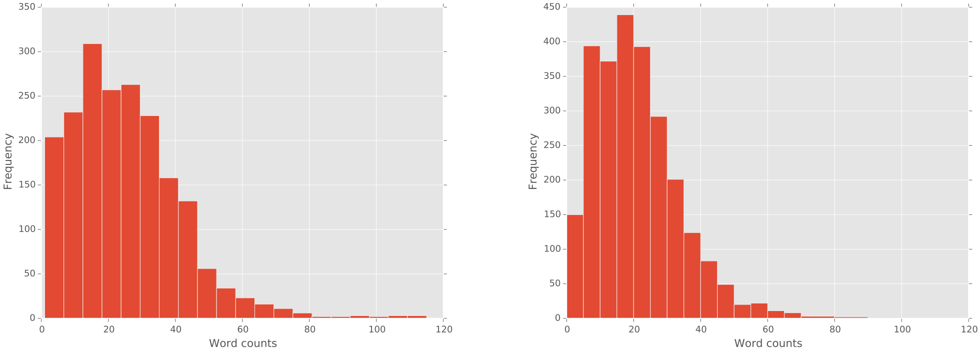
In this work we perform an empirical comparison between our rule-based system and machine learning approaches, tailored to the cosmetics domain for French language. The three systems use the same ensemble of basic linguistic features in order to obtain comparable systems and perform a fair comparison. We study the behavior of the systems in two different types of resources : 1) a well written corpus of journal articles, and 2) a corpus from a blog. In the following section (Section 2) we introduce the data used to develop and to evaluate our systems (described in Section 3). Finally, we compare the systems and discuss the results (Section 4).

2 Data

For the evaluation of the different systems two datasets were constructed. The first dataset, called MAG, consists of 99 articles (published between 6 of February 2012 and 30 of April 2012) from four French magazines on the domain of cosmetics : Beauté Infos, Féminin Pratique, Cosmétique Hebdo and Cosmétique Mag. It contains 1,944 sentences. The second dataset, called BLOG, contains 118 consecutive articles (published between 14 of June 2014 and 19 of November 2014) from the French blog `http://www.justsublime.fr`. It contains 2,613 sentences.

Dataset	**Sentences**	**Tokens**	Brand	Product	Division	Group	Range
MAG	1,944	46,029	748	164	103	250	61
BLOG	2,613	52,982	334	539	0	3	40

TABLE 1: Details of the datasets used in the comparison. Number of sentences and tokens as well as the partition of the five categories of named-entities in the three corpora.



(a) MAG

(b) BLOG

FIGURE 1: Word counts for the two datasets.

Both datasets were manually annotated by an expert according to five categories of named entities : product (*e.g. Shampooing Doux*), range (*e.g. Elsève*), brand (*e.g. L'Oréal Paris*), division (*e.g. L'Oréal Produits Grands Publics*), group (*e.g. L'Oréal*). Table 1 presents the properties for each annotated dataset. We highlight the fact that product and brand names are well represented categories in our datasets, while division, group, and range categories are less well represented. Figures 1a and 1b present the distributions of the number of words for each sentence in MAG and BLOG. Interestingly, the majority of the sentences contains less than 20 words and a lot of short sentences (under 10 words).

3 Methods and Setup

Our objective consists in recognizing 5 types of named entity : products, names of ranges, brands, divisions, and groups in the domain of Cosmetics. We compare one rule-based method relying on a syntactic analysis with two state-of-the-art machine learning approaches which we briefly describe in the following sections.

3.1 Rule-based method

Our first system, called Renco (*Reconnaissance d'Entités Nommées pour la COsmétique*), is based on a set of linguistic rules which use the output of a syntactic parser and a domain-specific dictionary (Lopez *et al.*, 2014). The training set for developing rules consisted of journalistic articles from French magazines (Beauté Infos, Féminin Pratique, Cosmétique Hebdo and Cosmétique Mag). Obviously, these articles are not included in MAG which is dedicated to the evaluation of the systems. Our linguistic rules can be classified into two categories :

— lexico-syntactic evidence rules (based on the well-known principles of (Hearst, 1992) and (McDonald, 1996)). Such rules are based on hyponymic and hyperonymic relations (for example in *parfum tel qu'Angel*). We also used internal contextual rules where internal evidence is a term included in an entity and that enables the annotation with a strong reliability. For example *Clarins Fragrance Group* where *Group* indicates clearly the type of this named entity.

— syntactic evidence rules consists of rules of coordination and hierarchical 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. Hierarchical rules are based on semantics of prepositions and enable to structure the extracted data (for example *Perfect Mousse est un produit de Schwarzkopf*).

Each rule outputs a score when it is triggered and a final score is calculated for each type of entity. The interested reader can refer, for a more detailed description of the system, to (Lopez *et al.*, 2014).

3.2 Machine learning methods

We employ two machine learning approaches. A CRF approach is a discriminative method which does not make any hypothesis over the data. CRF has exhibited state-of-the-art results in many NLP tasks also for NER in the e-Commerce domain (Putthividhya & Hu, 2011). The second method is a Learning to Search (L2S) approach which represents a family of algorithms for structured prediction tasks (Daumé III *et al.*, 2014). These methods decompose the problem in a search space with states, actions and policies and then learn a hypothesis controlling a policy over that state-action space. In this case each example in the training data is used as a reference policy (labels to be assigned at each token) and the learning algorithm tries to approximate it. This technique resembles reinforcement learning methods with the difference that in the latter one does not have a reference policy but discovers it through trial-and-error.

3.3 Setup

Since the rule-based system computes a final score for each candidate type of a given entity, a threshold has to be defined in order to decide which of the types will be used for the final annotation. This threshold has been empirically defined according to (Lopez *et al.*, 2014) and fixed to 0.80.

As for the machine learning methods an important step concerns the selection of features to be used by the algorithms. In the spirit of a fair comparison of the symbolic and the statistical systems, for both approaches, we try to use the same set of features as in the handcrafted rules : n-grams, part-of-speech tag, the lemma, the stemmed form as well as suffixes and prefixes of the target word and the words around it. We also use features indicating whether a word starts with a capital letter inside a sentence, whether the word is capitalized or not and also whether the word equals '&'. Additionally, we add two binary features when the previous word is "de" or "chez". Table 2 summarizes the features generated.

Both the rule-based and the machine learning systems use the Holmes syntactic parser (Bittar *et al.*, 2014) in order to extract linguistic features such as Part-Of-Speech tags, lemma, forms, and syntactic dependencies.

We use the Begin-In-Out (BIO) encoding in order to assign the labels and to transform the problem to a chunking task. We provide an example of the labeling on a snippet of the training data :

		Group								Brand		Brand		
Il	rejoint	Pierre	Fabre	en	2008	comme	directeur	des	marques	Ducray	et	A	-	Derma
O	O	B-G	I-G	O	O	O	O	O	O	B-B	O	B-B	I-B	I-B

Feature	Description
w_0,w_{-n},w_{+n}	Current token and window of -n :+n for n={1,2}
CAPS	All letters are capital
CapInitial	First letter is capital
Lemma(w),Stem(w)	Lemma and stem of current token
POS(w)	Part-of-speech tag of current token
w == &	Current token is &
w_{-1}==(de $\Vert$ chez)	Previous word is “de” or “chez”
affix(-3w :3w)	Prefixes/suffixes of words around current token

TABLE 2: Features used for the machine learning approaches.

As tools we use the Vowpal Wabbit [1] system which implements several L2S algorithms and the CRF++ toolkit [2] implementing L-BFGS and MIRA optimization for CRF. For tuning the hyper-parameters of each method we perform a random search coupled with cross-validation on the training set. For L2S we tuned the following parameters : learning_rate $\in [0.1, 4]$, passes $\in [2, 15]$, search_history $\in [1, 10]$. For CRF, the regularization parameter c $\in [0.0625, 16]$, max_iter $\in \{128, 256, 512\}$, and we used L-BFGS for learning.

Each dataset is split randomly in a training and a test set with 80% and 20% respectively. We evaluate the different systems solely in the test set using precision, recall and f-measure for each type of entity.

4 Results

Table 3 presents the results for the different systems across the two datasets in terms of precision (P), recall (R), and F-measure (F1). We omit the results for the non-entities for clarity of presentation.

Regarding MAG, it appears that the rule-based system obtains the best performance regardless the type of entities with F1 between 74.8 and 90.7 (standard deviation : 0.07). Both machine learning approaches obtain lower results. We observe that the CRF has a large variation in its performances : F1 goes from 20.9 to 81.9 (standard deviation : 0.27). Similarly, L2S results vary from 45.7 to 84.8 (standard deviation : 0.18). L2S and CRF gets a good precision but with a very low recall for products and ranges. Both learning methods outperform Renco for divisions that are under-represented in our corpus.

Concerning BLOG, all the systems obtain lower performances with this dataset certainly due to the lower quality in terms of grammar and structure implying a negative impact on the syntactic analysis. L2S obtains better results than Renco and CRF regarding brand names and ranges of product : respectively F1=0.80 and 0.76. However, Renco obtains the best results concerning recognition of

[1] http ://hunch.net/vw

[2] http ://crfpp.googlecode.com

		MAG			BLOG		
		Renco	**L2S**	**CRF**	**Renco**	**L2S**	**CRF**
Brand	Pr.	**93.19**	84.30	82.79	73.63	76.85	**83.13**
	R.	**88.30**	75.80	71.77	81.0	**83.0**	69.0
	F1.	**90.68**	79.83	76.88	77.14	**79.80**	75.40
Product	Pr.	**83.05**	78.78	64.28	**69.10**	68.42	63.52
	R.	**68.05**	36.11	12.50	**67.69**	60.0	55.38
	F1.	**74.80**	49.52	20.93	**68.39**	63.93	59.17
Division	Pr.	94.59	**95.45**	86.00	-	-	-
	R.	63.63	76.36	**78.18**	-	-	-
	F1.	76.08	**84.84**	81.90	-	-	-
Group	Pr.	**1.0**	70.45	92.18	-	-	-
	R.	**72.27**	70.45	67.04	-	-	-
	F1.	**87.17**	70.45	77.63	-	-	-
Range	Pr.	**1.0**	66.66	77.77	75.0	88.88	**1.0**
	R.	**65.21**	34.78	30.43	50.0	**66.66**	58.33
	F1.	**78.94**	45.71	43.75	60.0	**76.19**	73.68

TABLE 3: Results for all competing methods across the two datasets. Hyphenation is used when there are no predictions for the category or no tags were present in the dataset.

product names : F1=0.68. Renco, L2S, and CRF has a similar standard deviation (0.08) which can be explain by the fact that nor divisions neither groups were quantitatively representative in BLOG. Both L2S and CRF obtain better or comparable precision in the case of products with that of Renco but at the expense of lower recall.

Concluding, Renco showed a stable behavior in both corpora with lower performance in the BLOG dataset where L2S and CRF exhibit a more robust behavior. Note that the development set of Renco consists only of journalistic articles.

5 Conclusion and Further Work

In this work we presented an empirical comparison of a symbolic approach and machine-learning approaches for named entity recognition in the cosmetics domain. The different systems were compared on two datasets (journalistic articles and blog posts). The rule-based system outperformed both learning approaches in the journalistic corpus while the latter exhibited a more robust behavior in the blog corpus which contains noisy text.

For future work, we envisage to enhance the generation of features for the machine learning approaches by a) using dense distributed representations (Mikolov *et al.*, 2013) and b) transform the rules of Renco to features. We also intend to introduce an approach for learning automatically the weights of the rules inside LRB towards a hybrid system.

Références

BICK E. (2004). A named entity recognizer for danish. In *LREC*.

BITTAR A., DINI L., MAUREL S. & RUHLMANN M. (2014). The dangerous myth of the star system. In *LREC*, p. 2237–2241.

DAUMÉ III H., LANGFORD J. & ROSS S. (2014). Efficient programmable learning to search. *CoRR*, **abs/1406.1837**.

DÍAZ M. L. (2014). *L'onomastique des parfums : EN Presencia y renovación de la lingüística francesa*. Ediciones Universidad de Salamanca.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, p. 539–545 : Association for Computational Linguistics.

HEPP M. (2008). Goodrelations : An ontology for describing products and services offers on the web. In *Knowledge Engineering : Practice and Patterns*, p. 329–346. Springer.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, p. 282–289 : Morgan Kaufmann.

LOPEZ C., SEGOND F., HONDERMARCK O., CURTONI P. & DINI L. (2014). Generating a resource for products and brandnames recognition. application to the cosmetic domain. In *LREC*, p. 2559–2564.

MARRERO M., URBANO J., SÁNCHEZ-CUADRADO S., MORATO J. & GÓMEZ-BERBÍS J. M. (2013). Named entity recognition : fallacies, challenges and opportunities. *Computer Standards & Interfaces*, **35**(5), 482–489.

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

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

PUTTHIVIDHYA D. P. & HU J. (2011). Bootstrapped named entity recognition for product attribute extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 1557–1567.

RITTER A., CLARK S., ETZIONI O. *et al.* (2011). Named entity recognition in tweets : an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1524–1534 : Association for Computational Linguistics.

SEKINE S. & NOBATA C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In *LREC*, p. 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*, p. 418–423 : Association for Computational Linguistics.

ZHAO J. & LIU F. (2008). Product named entity recognition in chinese text. *Language Resources and Evaluation*, **42**(2), 197–217.