Recognizing Named Entities using Automatically Extracted Transduction Rules

Nouvel Damien, Jean-Yves Antoine, Nathalie Friburger, Arnaud Soulet

Recognizing Named Entities using Automatically Extracted Transduction Rules

Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger, Arnaud Soulet

Université François Rabelais Tours, Laboratoire d'Informatique
3, place Jean Jaures, 41000 Blois, FRANCE
{damien.nouvel, jean-yves.antoine, nathalie.friburger, arnaud.soulet}@univ-tours.fr

Abstract

Many evaluation campaigns have shown that knowledge-based and data-driven approaches remain equally competitive for Named Entity Recognition. Our research team has developed a symbolic system based on finite state tranducers, which achieved promising results during the Ester2 French-speaking evaluation campaign. Despite these encouraging results, manually extending the coverage of such a hand-crafted system is a difficult task. In this paper, we present results about the use of text mining techniques to automatically enrich our system's knowledge base. We exhaustively search for lexico-syntactic patterns, that recognize named entitites boundaries. We assess their efficiency by using such patterns in a standalone mode and in combination with the existing system.

1. Introduction

Named Entity (NE) Recognition (NER) is an information extraction process that aims at finding and categorizing specific entities (proper names, time expressions, amounts, etc.) in natural language streams. Those streams may be produced in various ways (e.g. electronic written documents, speech transcripts). The extracted entities may be used by higher-level tasks for different purposes, as Information Retrieval. Ester2, the latest evaluation campaign for NER over French oral corpora (more details on section 5.) has shown the good performances achieved by symbolic systems, often using transducers as a technical ground.

Our research team also developed such a NER system, based on Unitex[1] and transducer cascades, able to take into account diverse linguistic information (morphology, POS tagging, proper names lists, lexical evidences). The system relies on local grammars (Gross, 1997), describing NEs by taking advantage of clues that a human knows to be relevant. It has initially been designed to process written text. On French newspapers, it scores around 95% (f-measure) in recognizing persons, organizations and locations. More recently, it has been adapted to process oral transcripts (requiring more robust processings because of disfluencies and greater freedom in language-form), performing around 75% (f-measure). For the Ester2 campaign, three major issues were identified: coverage (finding all NE mentions, even those out-of-vocabulary), type disambiguation (especially for metonymies, which occur quite often in sport news, for instance), boundaries (much dependent on the campaign's annotation scheme).

Our goal is to investigate how rules, supposed to detect boundaries of categorized entities into data, may be discovered and parametrized in an automatic manner, and used as linguistic patterns to enrich our existing NER system. To this end, we use text mining techniques to extract from an annotated corpus transduction rules that independently detect beginning or ending boundaries of NEs. It releases the constraint for the system to match the whole entity, while still remaining in a transducer-like approach. Such transduction rules are then applied to recognize NEs. Preliminary experiments show that this standalone system successfully completes the symbolic system.

Section 2. presents and compares approaches for NER. In Section 3. and 4., we describe how lexico-syntactic sequential patterns may be extracted from annotated corpora and used as a standalone system. Next, Section 5. reports experimental results on French oral corpora.

2. Related Work

In the 90's and until now, several symbolic systems have been designed that, often, make intensive use of regular expressions formalism to describe NEs. Those systems often combined external and internal evidences (McDonald, 1996), as patterns describing contextual clues and lists of proper names by NE categories. Those systems achieve high accuracy, but, as stated by Mikheev et al. (1999), even when growing resources, coverage remains an issue.

Machine learning introduced new approaches to address NER. The problem is then stated as categorizing words that belong to a NE, taking into account various clues (features) whose weight is automatically parametrized using statistics from a *training* corpus. Among these methods, some only focus on the current word under examination (maximum entropy, SVM) (Borthwick et al., 1998), while others also evaluate stochastic dependencies (HMM, CRF) (McCallum and Li, 2003). Generally speaking, these approaches output the most probable sequence of labels for a given sentence. This is generally known as the "labeling problem", applied to NER.

A more detailed survey of NER classification techniques has been made by Nadeau and Sekine (2007). Most approaches rely on pre-processing steps that provide additional information about data, often Part-Of-speech (POS) tagging and proper names lists, to determine how to automatically *annotate* a text by detecting boundaries of NEs. We focus on those boundaries, as *beginning or ending markers* that the system intends to insert. Our main proposal is to extract patterns that are strongly correlated to

[1] http://www-igm.univ-mlv.fr/unitex/

$\mathcal{D}$	
Sent.	Patterns from $\mathcal{L}_I$
s_1	The american <pers> president Barack Obama </pers> has arrived in <loc> Moscou </loc>.
s_2	There he has seen the former <pers> chancelor Michelle Bachelet </pers>.
s_3	The <pers> president Dimitri Medvedev </pers> was not present on the beautiful <loc> square Vladimir Lenine </loc>.

Table 1: Sentences from an annotated corpus

the detection of one or many markers. That is to say, transducers are not constrained to necessarily recognize both boundaries of NEs. Thus, we separate the detection of markers denoting beginning and/or ending of NEs from the subsequent determination of actual NEs boundaries as two different tasks.

3. Mining Patterns from Corpus

3.1. Extracting Patterns

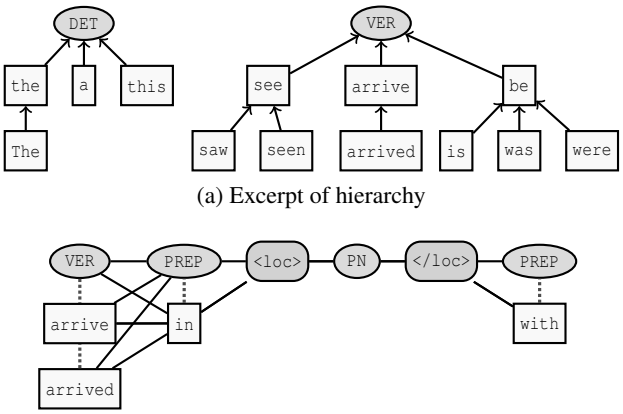
We use data mining techniques to process natural language. In this context, what is detected as a sentence will be considered as a sequence of items, precluding the extraction of patterns accross sentences. Two alphabets are defined: $\mathcal{W}$, words from natural language, and $\mathcal{M}$ as *markers*, the tags delimiting NEs categories (e.g. person, location, amount). The annotated corpus $\mathcal{D}$ is a multiset of sequences based on items from $\mathcal{W} \cup \mathcal{M}$. Table 1 exemplifies this with $\mathcal{W} = \{$`The`, `new`, `president`, ...$\}$ and $\mathcal{M} = \{$`<pers>`, `</pers>`, `<loc>`, `</loc>`, `<time>`, ..., `<org>`, ...$\}$.

The preprocessing step extends the language $\mathcal{W}$ to $\mathcal{W}^*$ by lemmatizing and applying a Part-Of-Speech (POS) tagger. This results in a hierarchy where each token may gradually be generalized to its lemma or POS. For instance in Table 1, the pattern language contains items $\{$`arrive`, `see`, `VER`, `JJ`, `DET`, `NN`, `PN` ...$\}$. The POS tagger distinguishes common nouns (`NN`) from proper names (`PN`). Furthermore, for proper names, we delete the token and the lemma, only keeping `PN` category, to avoid extracting patterns that would be specific to a given proper name. Figure 1 illustrates how POS categories are organized as a hierarchy and what patterns may be mined through an example of a sequence.

We exhaustively extract contiguous patterns over this language. For instance, in Figure 1, patterns such as '`VER in <loc> PN`' or '`PN </loc> with`' are extracted. The hierarchy and properties of sequential patterns allow to partially order them.

3.2. Filtering Patterns as Informative Rules

We mine a large annotated corpus to find generalized patterns that co-occur with NE markers. As usual in data mining, we set thresholds during extraction based on two interestingness measures: support and confidence. The *support* of a pattern P is its number of occurrences in $\mathcal{D}$, denoted by $supp(P, \mathcal{D})$. The greater the support of P, the more general the pattern P. Moreover, as we are only interrested in patterns correlated to markers, a rule R is defined



(a) Excerpt of hierarchy

(b) Possible paths to mine patterns for the example sentence "arrived in `<loc>` Moscou `</loc>`"

Figure 1: POS hierarchy and example of tagged sentence

as a pattern containing at least one marker. To estimate empirically how much R is accurate to detect markers, we calculate its *confidence*. A function $suppNoMark(R)$ returns the support of R when markers are omitted both in the rule and in the data. The confidence of R is:

$$conf(R, \mathcal{D}) = \frac{supp(R, \mathcal{D})}{suppNoMark(R, \mathcal{D})}$$

For instance, consider the rule $R =$ '`the JJ <pers> NN PN`' in Table 1. Its support is 2 (sentences s_1 and s_2). But its support without considering markers is 3, since sentence s_3 matches the rule when markers (`<pers>` in rule and `<loc>` in "the beautiful `<loc>` square Vladimir") are forgotten. Thus the confidence of R is only 2/3.

The whole collection of transduction rules exceeding a minimal support and confidence thresholds is used as a knowledge-base. In practice, the number of discovered rules remains very large (especially when minimal support threshold is low). Thus, we decide to filter-out the redundant rules. We consider two rules to be redundant if they have same support and confidence, and furthermore if one is a generalization of the other. Over a set of redundant rules, we only select the most specific ones, named *NR-rules*.

4. NER using Informative Rules

The difficulty we are now facing is to determine whenever a transduction rule should be applied to insert a marker in an unseen text. To tackle this problem, the marking algorithm described in Section 4.2. selects the most probable annotation according to the probability model defined by Section 4.1..

4.1. Probability Model

As previously mentioned, instead of assigning a category to words (or tokens), we use transduction rules to insert markers at diverse positions in the sentence. At any position, we have the choice between adding beginning or ending markers for NE categories (e.g. `<cat>` or `</cat>`) or not to do so, what we denote by inserting a 'void marker"

($\emptyset$). As a first approximation, we make the assumption that the presence (or absence) of markers at different positions within a sentence are mutually independent.

Thus, we represent local probabilities for inserting any markers or no marker $\{\emptyset, \texttt{<pers>}, \texttt{</pers>}, \texttt{<loc>}, \dots\}$ at a given position i as a random variable $P(Mi = m_{j_i})$. This probability depends on the set of rules that have been triggered at the current position $R_1, R_2 \dots R_k$. We compute those probabilities using a Maximum Entropy (MaxEnt) classifier [2] for which rules are considered as features. The classifier is learned (parametrized) using training data where markers and triggered rules are known. Consequently, probabilities of markers may be retrieved from the model at any position of the sentence depending on what rules have been triggered at that given position. Finally, we use those local probabilities to approximate the probability of making n decisions (NE or void markers) over a sentence as:

$$P(M_1 = m_{j_1}, M_2 = m_{j_2}, \dots, M_n = m_{j_n}) \approx \prod_{i=1\dots n} P(M_i = m_{j_i})$$

What would be considered as the most probable affectation of markers within possible ones would maximize that measure.

4.2. Marking Algorithm

Whatever markers are the most probables, the resulting annotation has to be *valid* according to an annotation scheme. For our purpose, we need a flat (no imbrication) xml-like annotation. From our point of view, we consider adding a marker as making a transition (for instance, a `<loc>` marker moves from a "no-NE" state to a "loc" state, and afterwards only $\emptyset$ or `</loc>` could be inserted). It is straightforward to see that to determine the most probable annotation at a given point of the sequence, we have to compute the most probable paths to possible states.

We implemented an algorithm based on dynamic programming techniques. It only keeps in memory N+1 hypotheses, where N is the number of considered NE categories. Most probable annotation path and its probability are stored for each possible NE "state" as markers are encountered. For this purpose, at any position of the sentence, for any states, are confronted the previous states probabilities combined with the transition probabilities. For instance, the probability of the "loc" state is set as the path of highest probability between the preceding "loc" hypothesis where the probability of $\emptyset$ is taken into account and the "no-NE" state where the marker `<loc>` is inserted. As the sequence is examined, only most probable annotations are retained per state, and the resulting annotation, when the whole sentence has been processed, is the "no-NE" hypothesis.

5. Experimental Results

5.1. Data: French Radio Transcripts

Our work is dedicated to the recognition of NEs in French oral transcripts. This task is more challenging on this kind of data: the data is noisy (hesitations, repairs, speech turns) and sentence boundaries are harder to detect. This accordingly lowers performance of POS tagging and, at a higher level, requires a more robust approach to find entities.

Corpus	Tokens	Sentences	NEs
Ester2-corr	40 167	1 300	2 798
Ester2-held	48 143	1 683	3 074

Table 2: Characteristics of Ester2 corpora

The French Ester2 evaluation campaign included NER on transcribed texts (Galliano et al., 2009). The competing systems had to recognize persons, locations, organizations, products, amounts, time and positions. Entitites were manually annotated for evaluation purposes. As reported by Nouvel et al. (2010), annotation inconsistencies have been pointed out, the corpus was re-annotated more consistantly for a first half (Ester2-corr), while the second part was held out (Ester2-held). Detailed characteristics of those corpora are presented in Table 2.

5.2. Informative Rules Extraction Results

Patterns will be extracted over the Ester2-corr corpus. The POS processing was made using TreeTagger (Schmid, 1994), which uses decision trees to robustly tokenize and simultaneously tag tokens with POS categories and lemmatize words (on French written texts, this tool provides high accuracy, more than 90% but, as far as we know, no evaluation has been made over oral transcriptions). The mining task requires many optimizations and we used a level-wise algorithm (Mannila et al., 1997) which leverages the generalization over patterns to mine frequent ones. Table 3 reports the number of rules, the number of non-redundant rules and the gain (i.e., the ratio between the number of rules and that of non-redundant ones). This elimination of redundant rules leads to a significant reduction (36 at low frequency) without loss of information that facilitates the use of this collection as knowledge-base.

Corpus	Sup.	Conf.	Rules	NR-rules	Gain
Ester2-corr	10	.5	2 270	1 119	**2.03**
	5	.5	28 047	3 673	**7.63**
	3	.3	458 875	12 653	**36.27**

Table 3: Extraction over Ester2 corpus at support and confidence thresholds

5.3. Standalone System

We use Ester2-corr in a 12 folds cross validation to extract rules (11 files) and to evaluate accuracy of the predicted markers (12th file). Training the MaxEnt classifier using extracted rules as features requires a separate corpus: we merged Ester2-held with another corpus (Eslo) containing similar annotations for this purpose. In order to retrieve a set of rules that covers as much as possible actual markers in texts, we hereby extract rules at low support (3) and confidence (0.3) thresholds. With this exhaustive set of rules, only 52 markers out of 5196 (1%) are undetectable by the

[2] http://homepages.inf.ed.ac.uk/lzhang10/maxent.html

Actual markers	Predicted markers															
		tot	∅	<pers>	</pers>	<loc>	</loc>	<org>	</org>	<fonc>	</fonc>	<time>	</time>	<amo>	</amo>	rec.
	∅	**27803**	**27168**	46	5	114	68	91	75	28	28	77	76	14	13	0.98
	<pers>	**583**	86	**430**		20	1	26	1		18	1				0.74
	</pers>	**592**	48		**470**		45		27			1	1			0.79
	<loc>	**700**	162	20	2	**394**		114	1		2	3	2			0.56
	</loc>	**698**	137	2	16	2	**407**		*127*			4	3			0.58
	<org>	**448**	203	*30*		*45*		**157**		2	6	3	2			*0.35*
	</org>	**443**	176		*59*		*69*		**122**		2	5	8	2		*0.27*
	<fonc>	**225**	84	1	2	3		2		**129**		4				0.57
	</fonc>	**219**	112	*27*	6		*10*		14		**48**	1	1			*0.22*
	<time>	**508**	249	2	4	1	12		4			**223**	4	8	1	0.44
	</time>	**507**	200	1			6		2			2	**293**	1	12	0.57
	<amo>	**130**	98			1	1					*6*	*2*	**21**	1	*0.16*
	</amo>	**133**	79		1				1				*17*		**35**	*0.26*
	prec.		0.94	0.77	0.83	0.68	0.66	0.40	*0.33*	0.81	0.46	0.68	0.72	0.46	0.56	

Table 4: Confusion matrix between rule markers using a MaxEnt classifier (amo = amount)

model because no rules are triggered at the considered position.

Table 4 shows the results of this experiment as a confusion matrix. At each position, we compare the most probable marker obtained by applying the maximum entropy model (itself relying on locally triggered rules) to the actual marker. Indeed, the accuracy of *not* introducing a marker is quite high. The person markers seem well detected. The most difficult markers are those of organizations, which have both low precision and recall and reveal a high rate of ambiguities. For the MaxEnt classifier, a high ambiguity prevents from inserting any marker.

Table 5 gives more details about kind of errors on NEs: insertions, deletions, types, extents. Those are weighted to compute the Slot Error Rate (SER) (Makhoul et al., 1994). Results show that, for any support threshold, the MaxEnt model allows to extract rules at low confidence, that is to say, even very generic (and thus less confident) rules may be included in the MaxEnt model. Globaly, using those rules as a standalone system remains unsufficient regarding performance compared to state-of-the-art, but opens up great possibilities for coupling.

Sup.	Conf.	Ins.	Del.	Typ.	Ext.	SER
3	0.3	112	669	192	228	**45,54**
3	0.5	97	913	178	233	56,93
3	0.7	44	1657	192	164	84,4
5	0.3	92	681	202	241	**44,77**
5	0.5	97	1222	205	194	69,81
5	0.7	50	1526	184	172	78,48
10	0.3	89	748	205	222	**47,13**
10	0.5	68	855	194	197	49,59
10	0.7	38	1211	162	159	61,11

Table 5: Detailed results

5.4. Coupling Rules with a Symbolic System

We aim at improving performances of an existing system with extracted patterns. Our symbolic system achieves is precise, but it lacks coverage because it would have to describe all regular expressions that may constitute a NE. Our idea is that automatically extracted patterns may recover some of the NE omitted by the symbolic system to improve its performance. We chain systems, where the annotation produced by the symbolic system is considered as certain: the marking algorithm will only try to add NE where the symbolic system did not find any NE.

Table 6 reports the initial symbolic system's results, the differences of errors by NE categories and the resulting coupled system's performance. The symbolic system alone outperforms (29 SER) our standalone system using rules (44,77 SER). By coupling systems, we observe a significant improvement of the symbolic system's output, from 29 (hybrid) to 27.7 (coupled) SER. The insertion of a relatively small amount (5) of false-positive (Ins. total) is the counterpart for the detection of 51 NEs omitted by the symbolic system. Mainly, the transduction rules allows to recover organizations, persons and locations.

	Ins.	Del.	Typ.	Ext.	SER
Symbolic	43	348	171	257	29.0
fonc	0	-1	+1	0	28.8
loc	+4	-15	+3	+1	16.8
org	0	-13	+11	0	52.8
pers	+1	-20	0	+8	15.3
time	0	-2	0	0	24.6
total	+5	-51	+19	+8	-1.3
Coupled	48	297	190	265	27.7

Table 6: Error differences on CasEN with extracted rules

We also isolated and manually examined rules that were responsible for the decrease of deletion errors (coverage). Most of these rules are short and generalized rules, and quite frequently inserting only one marker (for instance 'from <pers> PN PN' or 'to <loc> PN'). Interestingly, two time expressions have been found thanks to the separate detection of the beginning and the ending markers using local clues: 'for <time>' and 'years </time>' (recognizing "for a few years" for instance). How those shallow rules may be taken into account by the knowledge base of the symbolic system remains to be investigated.

6. Conclusion

In this paper, we reported experimentations on the use of text mining techniques to automatically enrich a knowledge-based NER system. We implemented a prototype which extracts patterns correlated to NE markers. The system exhaustively looks for transduction rules from an annotated training corpus and filters out those of interest. During the mining process, the text is represented as a sequence of items, which may be generalized using a hierarchy through POS categories, and where the beginning or ending markers of NEs may be separately mined.

The quality of patterns and their potential to recognize entites has been assessed and allowed us to state which are the most efficient and what markers categories remain to be improved. These experiments also investigated the idea of separately evaluating the probability to begin or end an entity, an algorithm being afterwards responsible for finding a valid and probable annotation. The resulting system was used as a postprocessing behind a symbolic system, showing significant improvement of the performance. This work provides us with some interesting directions for improving a symbolic NER system, including in its foundations.

References

Borthwick, Andrew, John Sterling, Eugene Agichtein, and Ralph Grishman, 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *6th Workshop on Very Large Corpora (WVLC'1998)*.

Galliano, Sylvain, Guillaume Gravier, and Laura Chaubard, 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *10th Conference of the International Speech Communication Association (INTERSPEECH'2009)*.

Gross, Maurice, 1997. The construction of local grammars. *Finite-State Language Processing*:329–354.

Makhoul, John, Francis Kubala, Richard Schwartz, and Ralph Weischedel, 1994. Performance measures for information extraction. *DARPA Broadcast News Workshop*:249–252.

Mannila, Heikki, Hannu Toivonen, and A. Inkeri Verkamo, 1997. Discovery of frequent episodes in event sequences. In *Data Mining and Knowledge Discovery (DMKD)*, volume 1.

McCallum, Andrew and Wei Li, 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *13th Conference on Computational Natural Language Learning (CONLL'2003)*.

McDonald, David D., 1996. Internal and external evidence in the identification and semantic categorization of proper names. *Corpus Processing for Lexical Acquisition*:21–39.

Mikheev, Andrei, Marc Moens, and Claire Grover, 1999. Named entity recognition without gazetteers. In *9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'1999)*.

Nadeau, David and Satoshi Sekine, 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30:3–26.

Nouvel, Damien, Jean-Yves Antoine, Nathalie Friburger, and Denis Maurel, 2010. An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign. In *7th International Language Resources and Evaluation (LREC'2010)*.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In *2nd International Conference on New Methods in Language Processing (NEMLP'1994)*.