



# Efficient combined approach for named entity recognition in spoken language

Azeddine Zidouni<sup>1</sup>, Sophie Rosset<sup>2</sup>, Hervé Glotin<sup>3</sup>

<sup>1</sup>LSIS-CNRS, Marseille, France.

<sup>2</sup>LIMSI-CNRS, Orsay, France.

<sup>3</sup>LSIS-CNRS, Toulon, France.

azeddine.zidouni@lsis.org, sophie.rosset@limsi.fr, glotin@univ-tln.fr

## Abstract

We focus in this paper on the named entity recognition task in spoken data. The proposed approach investigates the use of various contexts of the words to improve recognition. Experimental results carried out on speech data from French broadcast news, using conditional random fields (CRF) show that the use of semantic information, generated using symbolic analyzer outperform the classical approach in reference transcriptions, and it is more robust in automatic speech recognition (ASR) output.

**Index Terms:** Named entity recognition, automatic speech recognition.

## 1. Introduction

Traditional Named Entity Recognition (NER) is a task in which proper nouns and numerical information are extracted from documents and are classified into categories such as person, organization, and date. It is a key technology of Information Extraction (IE) and Open-Domain Question Answering [1]. It is also a fundamental component of a number of other language processing applications such as text clustering, topic detection, and summarization. While significant progress has been reported on the NER task, most of the previous approaches have generally focused on clean textual data [2]. Some have focused on speech data [3, 4, 5]. For extracting NE from spoken language, we use automatic speech recognition (ASR) results. However, applying NER to ASR results involves additional errors, which are caused by out-of-vocabulary words and discordance of acoustic/language models. Although continuous efforts to improve ASR itself are needed, developing a robust NER for noisy word sequences containing ASR errors is also important.

Different initiatives such as MUC [6] and TIPSTER [7] paved the way for the development of many current IE systems. In a short amount of time, systems were able to recognize named entities with precision and recall scores in the 90th percentile in narrow domains such as newswires about terrorist attacks. Work went on with the ACE [8] project which kept an equivalent definition of named entities.

The rest of paper is structured as follows. Section 2 provides an overview of symbolic and statistic approaches used in NER task. We describe in Section 3 the NER task definition in ESTER 2 campaign. We present in Section 4 two NER methods: the LISMI symbolic analyzer and the LSIS annotator. We then propose a fusion protocol to improve NER results in both manual and ASR transcribed speech. In Section 5, we give experimental results and compare our approach to ESTER 2 participating sites. We conclude in Section 7.

## 2. Related work

Several approaches try to reduce the gap between the reference and predicted annotations by reducing ambiguities, they arise to improve robustness and portability. Mostly machine learning (ML) approaches are used in NER task like Support Vector Machine (SVM) [9], Decision Tree [10] or Hidden Markov Models [11]. More recently, Conditional Random Fields (CRF) [12] have been used in NER [13]. The main advantage of CRF is their flexibility to include a variety of features.

Most of the symbolic approaches to NER task rely on morphosyntactic and/or syntactic analysis as in [14]. Rule-based approaches are based on regular expressions or linguistic pattern and make use of dictionaries for recognizing named-entities. Rule-based approaches have proven to be quite successful, specifically when enhanced with syntactic analysis. In the ESTER 2 campaign [5], on manual transcriptions, the two systems based on syntactic analysis in addition to rules obtained the best results.

## 3. Task description

Our experiments have been done within the framework of the French Rich Transcription Program of Broadcast News ESTER 2 [5], which includes a NER task.

### 3.1. Corpus description

The provided corpus consists of 100 hours made from four French speaking radio channels manually transcribed. Each radio program is represented by a transcription file. The different NE defined in this task are: *location*, *organization*, *time*, *amount*, *function*, *person* and *production*. Each of them is split into a number of sub-categories. The chosen NE tagset is then made of 7 main categories and over 30 sub-categories:

- **person:** human person and fiction person,
- **location:** natural and geographical location, administrative district, road, mail address, fax and phone numbers, email, human construction,
- **organization:** political organization, education, commercial, non commercial, media and entertainment, location acting like an organization,
- **time:** date (absolute and relative) and hour,
- **amount:** 10 different categories have been defined covering a broad spectrum, from age to speed,
- **function:** 5 types of functions have been defined, military, political, administrative, religious and aristocratic,
- **production:** 4 different categories have been defined, vehicle, award, art and official documentation (law etc.).

Hierarchical definitions are possible. For example the tag *pers.hum* (human person) can include names and function which can include location or organization. For example:  $\langle pers.hum \rangle \langle fonc.pol \rangle$  président  $\langle /fonc.pol \rangle \langle pers.hum \rangle$  Amadou Toumani Touré  $\langle /pers.hum \rangle \langle /pers.hum \rangle$ . Only the seven top-level entities are evaluated.

### 3.2. Metrics used

To measure the performance of each model we used four evaluation measures: the recall (R), precision (P), F-measure (F), and the slot error rate (SER) [15]. The recall is the percentage of reference slots for which the hypothesis is correct. The precision is the percentage of slots in the hypothesis that are correct. The F-measure is defined as the weighted harmonic mean of recall and precision. Finally, the SER is more accurate and penalizing than F-measure. These measures are the same used in the ESTER 2 evaluation campaign.

## 4. Approaches used

The main objective of the work presented in this paper is to investigate which context of the words can perform NER task in transcribed speech. In the following sections, we present the LIMSI rule-based approach and the LSIS CRF-based approach. Then we present the third approach based on the combination of the both.

### 4.1. The LIMSI multi-level analyzer

The analyzer is part of the Ritel system [16], a spoken language dialog system in open domain which includes a question answering system. The analysis is *non-contextual* because each sentence (or turn or speech segment) is processed in isolation. The general objective of this analysis is to find the bits of information that may be of use for search and extraction, which we call *pertinent information chunks*. These can be of different categories: named entities, linguistic entities (e.g. verbs, prepositions), or specific entities (e.g. scores). The entity definition on which the system is based is hierarchical.

This system is rule-based and used *WMatch* [17] a tool developed at LIMSI. This engine matches (and substitutes) regular expressions using words as the base unit instead of characters. This property allows enables the use of classes (lists of words) and macros (sub-expressions in-line in a larger expression). *WMatch* includes also NLP-oriented features like strategies for prioritizing rule application, recursive substitution modes, word tagging (for tags like noun, verb, etc.), word categories (number, acronym, proper name, etc.). Analysis is multi-pass, and subsequent rule applications operate on the results of previous rule applications which can be enriched or modified. The unified nature of the analysis representation precludes us from using standard evaluation set to measure its performance. Still, some interesting quantitative results can be obtained in the context of question-answering. Specifically, using the QAsT 2008 data [18], we checked whether the reference answers had been correctly detected and typed. The results show that for French the correct detection rate is 87.4% in manual transcriptions. On automatic transcriptions, for French, the correct detection rate is 86.2% with a 11.0% of word error rate (WER), 86.2% with a 23.9% of WER and 78.2% with 35.4% of WER.

The Ritel analyzer provides more than 300 different types of words and multi-word expression, most of them being semantic and close to some NE definition. Obviously, because

this system is used within a question-answering system and a spoken dialog system, the need is different from a NE task. For the evaluation, we decided to adapt the Ritel analyzer to the ESTER 2 task. The adaptation is needed because of the differences in the entity definitions. Some entities have a direct mapping as shown in the Figure 1 while others are more complex as shown in Figure 2.

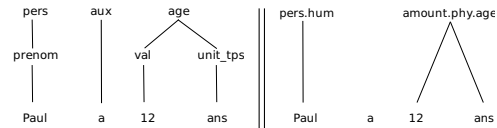


Figure 1: Comparison between Ritel (left) and ESTER 2 (right): case of **age**.

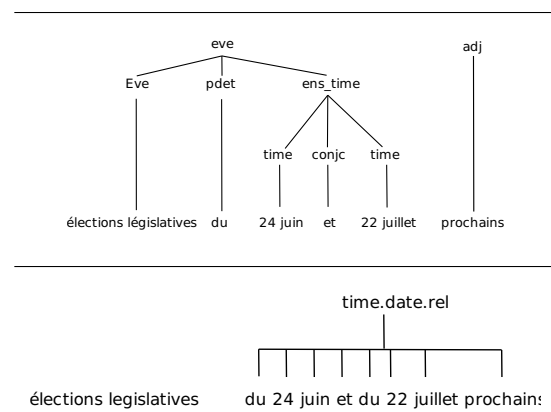


Figure 2: Comparison between Ritel (top) and ESTER 2 (bottom): case of **event** vs. **time**.

We developed an aligner which aligns the development data set to the same documents annotated by the Ritel system. This aligner allowed to automatically extract patterns which were then reintroduced in new analysis passes to remap the Ritel results to the types expected by the evaluation (LIMSI system).

### 4.2. The LSIS NER annotator

The power of graphical models lies in their ability to model many variables that are independent by a product of local functions that each depends on only a small number of variables. The ML models proposed in this paper are implemented using CRF approach. CRF are discriminative undirected graphical models which are conditionally trained to discriminate the correct sequence from all other candidate sequences without making independence assumption for features. The CRF approach allows the use of arbitrary, dependent features and joint inference over entire sequence (incorporate many features of the words in a single unified model) [19]. CRF approach constructs models which characterize input data. Each token of the input data can be defined with several features (also called attributes). These attributes can be used in the learning phase to improve prediction models.

Part of Speech (POS) tagging is the task of labeling each word in a sentence with its appropriate part of speech category. In the NER process we can include the POS and lemma annotations as word attributes (Figure 3). To perform the POS tagging

we use the *TreeTagger* [20]. We associate for each word  $w_i$  its POS category  $S(w_i)$  where  $S(w_i) \in \{NAM, VERB, DET, PRE, ADV, ADJ, KON, NUM, PUN\}$ . For example, the syntactic representation of the sentence  $\langle Albert Einstein was born on March 14, 1879 \rangle$  is  $\langle NAM NAM VERB VER PRE NAM NUM PUN NUM \rangle$ .

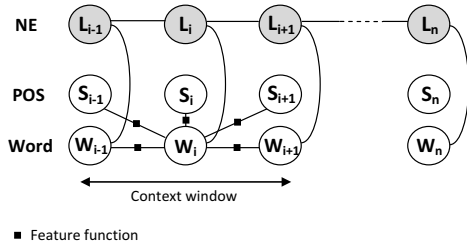


Figure 3: Example of CRF model using POS tags as attributes of the words and a context window of  $[-1, +1]$ .

To make the boundary of NE, we use the annotations in BIO format, which assigns to each token one of the labels:  $B$  – *Category* if the token begins a NE of type *Category*,  $I$  – *Category* if the token is inside a NE of type *Category*, and  $O$  if the token does not belong to any known NE type.

### 4.3. Fusion

To perform the annotation, each word  $w_i$  in CRF model can be represented by several families of features. Traditionally, atomic observations of the word are used (such as the word itself, capitalization, prefixes and suffixes, neighboring words, etc.). Additional information like syntactic and semantic features can be used to enrich the model. The fusion approach consists of using the Ritel system outputs as an input attributes in CRF model. We use the three lower levels of the Ritel hierarchical output as a CRF attributes instead of POS tags. Thus, to predict the NE tag of the word, in addition to the word, we use the annotations produced by the Ritel analyzer.

## 5. Experiments and results

The ESTER 2 corpus is divided into three parts, the *training* part (84%) which is used to train the models, the *development* (8%) which is used to adjust various parameters and the *test* (8%) which is used for the evaluations. The training and development data have been provided with manual transcriptions only whereas the evaluation has been done on manual transcripts and three different automatic transcriptions [5]. In these paper we present results on the manual transcriptions ( $test_{REF}$ ) and the first ASR output ( $test_{ASR}$ ) which is case-sensitive and obtained a WER of 12.11%.

The CRF-based models have been trained on the ESTER 2 training data. There were some differences between the training data (annotated in a first phase) and the development data (annotated in a second phase along with the test data). For example, some NE definitions have been extended and differences in frontier as in  $\hat{a} \langle time.rel \rangle demain \langle time.rel \rangle$  versus  $\langle time.rel \rangle \hat{a} demain \langle time.rel \rangle$  were found. We developed a rule-based aligner which built adaptation rules based on differences between the training and the development data. This system has been used as post-processing in the  $CRF_{Ritel}$  and  $CRF_{POS}$  systems. The last one is the LSIS system with this post-processing step.

To build the CRF models, an immediate context window of 2 token to the left and 2 token to the right of the current token  $[-2, +2]$  has been used. The  $CRF_{Ritel}$  system uses the output of the Ritel analyzer as input features in CRF model.

One difference between the manual transcriptions and the ASR output is the lack of punctuations in the ASR output. The NER systems presented in this paper work on the sentence level and assume the presence of standard punctuation, at least the periods and commas. A simple sentence-segmentation model has been trained on the training data using CRF and word-context as feature (is this word followed by a punctuation mark?). This model has been used to segment the ASR output in the  $test_{ASR}$  experiment.

Table 1: Slot Error Rate (SER) values for each NE type for the different systems.

NEs	LIMSI	LSIS	$CRF_{POS}$	$CRF_{Ritel}$
PERS	13.05	28.33	13.42	<b>12.30</b>
LOC	33.88	28.02	13.42	<b>9.08</b>
ORG	48.50	48.81	31.93	<b>27.46</b>
FUNC	47.30	65.31	33.92	<b>30.24</b>
PROD	94.66	73.97	69.82	<b>71.58</b>
TIME	17.70	25.65	26.47	<b>25.03</b>
AMOUNT	33.81	35.73	35.15	<b>33.10</b>
<b>Overall</b>	30.88	34.98	23.16	<b>20.26</b>

Table 1 shows the global Slot Error Rate (SER) and the SER for each NE type. The first observation is that the post-processing (introduced in  $CRF_{POS}$  and  $CRF_{Ritel}$ ) allows a major improvement in the results. Indeed, many errors were caused by the differences in *training* and the *dev/test* data. A decrease of the SER is observed on almost all the types (from  $-0.58\%$  to  $-31.39\%$ ). We can also notice that the use of semantic context improves the NE detection, especially in LOC ( $-24.80\%$  of SER), ORG ( $-21.04\%$ ), and FUNC ( $-17.06\%$ ) NE types compared to the CRF baseline approach (the LSIS system). PROD quality degrades with  $CRF_{Ritel}$  because there is not enough elements in the corpus to fully optimize its parameters. PERS and TIME NE types have a small improvement compared to results in  $CRF_{POS}$ , the POS tags gives a good informative components almost semantic informations.

Table 2: Comparison of NER task overall performance between our approach and ESTER 2 participating sites in manual ( $test_{REF}$ ) and automatic ( $test_{ASR}$ ) transcriptions [5].

%WER	$test_{REF}$			$test_{ASR}$		
	%SER	%Pre	%Rap	%SER	%Pre	%Rap
LIA	23.91	86.46	71.85	43.61	79.52	59.45
LIMSI	30.88	81.15	70.94	45.34	75.13	62.33
LINA	37.15	80.75	55.48	53.97	71.98	44.01
LI Tours	33.74	79.39	65.82	50.71	71.36	54.16
LSIS	34.98	82.65	73.07	49.03	72.81	60.98
Synapse	9.93	93.02	89.37	44.86	76.39	67.16
Xerox	9.80	93.61	91.50	44.60	58.91	70.06
$CRF_{Ritel}$	20.26	86.96	76.21	44.35	73.64	64.60

As shown in table 2, the proposed approach achieved the 3rd best SER in manual transcriptions, 20.26%, compared to the official results. The improvement is of 14.72% compared to

the LSIS result and of 10.62% compared to the LIMSI results. On ASR output, this new approach is more robust than the two best systems.  $CRF_{Ritel}$  presents a slight improvement (0.25%) over the best system. However, we can notice that there is very few difference between the best results. That observation tend to show that NER on ASR output, even with a low WER, is a problem and that the best systems on manual transcripts are not necessarily the more robust to handle speech recognition errors.

## 6. Discussion

The results (in table 1) show that some classes are more sensitive than others to the addition of semantic features. The types FUNC and ORG show a great improvement in the results. The semantic analysis provided by the Ritel system seems to give useful information (mainly triggers we suppose) to detect these types. Nevertheless, the confusion between ambiguous classes (ORG vs LOC) is still high even if lower with this new system (table 3). To decide if *Paris* is an organization or a localization is not only a question of semantic information but of lexical and syntactic information as observed in [21].

Table 3: Confusion table for the  $CRF_{POS}$  (top) and  $CRF_{Ritel}$  models (bottom).

%	PERS	LOC	ORG	FUNC	PROD	TIME	AMNT
PERS	92	6	< 1	1	0	0	0
LOC	2	91	5	0	0	< 1	0
ORG	7	18	73	< 1	< 1	1	0
FUNC	7	2	4	86	0	0	0
PROD	30	16	21	0	30	3	0
TIME	1	< 1	< 1	0	0	93	3
AMNT	< 1	0	0	0	0	8	90

%	PERS	LOC	ORG	FUNC	PROD	TIME	AMNT
PERS	94	3	< 1	1	0	0	0
LOC	2	94	3	0	0	< 1	0
ORG	4	17	76	< 1	< 1	1	0
FUNC	6	1	2	90	0	0	0
PROD	23	19	12	0	38	5	1
TIME	1	< 1	< 1	0	0	94	2
AMNT	< 1	0	0	0	0	7	93

## 7. Conclusion and Perspectives

In this paper we have presented a NER system that combines symbolic analyzer outputs with discriminative approach. We have shown that the use of the outputs of another system as a priori knowledge improves general quality of the annotations and makes the prediction more robust in the case of ASR output. In experiments using CRF the proposed approach showed a better SER results compared to classical approaches (LSIS and LIMSI).

Our approach can be applied to other tasks of spoken language processing. Our future work will focus on optimizing the contexts length for each NE type. The approach can also be extended to other semantic annotations of noisy inputs and introduce more speech-specific features such as harmonic to noise ratios (HNR) values.

## 8. Acknowledgements

This work has been partially financed by OSEO under the Quaero program.

## 9. References

[1] H. M. Voorhees and D. Harman, "Overview of the ninth text retrieval conference," in *(TREC-9)*, 2000.

[2] E. T. K. Sang and F. D. Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *In Proc. of CoNLL*, 2003.

[3] F. Kubala, R. Schwartz, R. Stone, and R. Weischede, "Named entity extraction from speech," in *In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[4] D. Palmer, J. D. Burger, and M. Ostendorf, "Information extraction from broadcast news speech data," in *In Proc. of the DARPA Broadcast News Workshop*, 1999.

[5] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *In Interspeech 2009*, 2009.

[6] SAIC, "Proceedings of the seventh message understanding conference (muc-7)," Tech. Rep., 1998. [Online]. Available: [http://www.itl.nist.gov/iaui/894.02/related\\_projects/muc/](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/)

[7] *Advanced Research Projects Agency. Proceedings of the TIPSTER Text Program (Phase II)*. Morgan Kaufmann, California, 1996.

[8] *Entity Detection and Tracking, Phase 1, ACE Pilot Study Task Definition*, 2000. [Online]. Available: <http://www.nist.gov/speech/tests/ace/phase1/doc/summary-v01.htm>

[9] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," in *Proceedings of COLING'02*, 2002, pp. 390–396.

[10] S. Sekine, "Description of the japanese ne system used for met-2," in *Proceedings of the 7th Message Understanding Conference*, 1998, [www.itl.nist.gov/iaui/894.02/related\\_projects/muc/](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/).

[11] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, and R. Weischedel, "Bbn description of the sift system as used for MUC7," in *Proceedings of the 7th Message Understanding Conference*, 1998.

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[13] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 188–191.

[14] C. Brun and C. Hagège, "Intertwining deep syntactic processing and named entity detection," in *EsTAL*, 2004, pp. 195–206.

[15] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *In Proceedings of DARPA Broadcast News Workshop*, 1999, pp. 249–252.

[16] B. van Schooten, S. Rosset, O. Galibert, A. Max, R. op den Akker, and G. Illouz, "Handling speech input in the ritel qa dialogue system," in *InterSpeech'07*, Anvers, Belgique, 2007.

[17] O. Galibert, "Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert," Ph.D. dissertation, Université Paris-Sud 11, Orsay, France, 2009.

[18] J. Turmo, P. Comas, S. Rosset, L. Lamel, N. Moreau, and D. Mostefa, "Overview of qast 2008," in *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September 2008.

[19] A. Zidouni and H. Glotin, "Semantic annotation of transcribed audio broadcast news using contextual features in graphical discriminative models," in A. Gelbukh (Ed.): *CICLing 2010, LNCS 6008*, pp. 279–290. Springer, Heidelberg (2010), March 2010.

[20] S. Helmut, "Part-of-speech tagging using decision trees," in *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[21] C. Brun, M. Ehrmann, and G. Jacquet, "A hybrid system for named entity metonymy resolution," pp. 118–130, 2009.