# Extended Semantic Tagging for Entity Extraction

**Narjès Boufaden*, Yoshua Bengio†, Guy Lapalme***

*Laboratoire RALI - Université de Montral
Québec, Canada
{boufaden,lapalme}@iro.umontreal.ca
†Laboratoire LISA - Université de Montral
Québec, Canada
bengioy@iro.umontreal.ca

**Abstract**

We present results of a statistical method we developped for the detection of what we define as generalized named entities from manually transcribed conversations. This work is part of an ongoing project for an information extraction system in the field of maritime Search And Rescue (SAR). Our purpose is to automatically detect relevant words and annotate them with concepts from a SAR ontology. Our approach combines similarity score vectors and topical information. Similarity vectors are generated using a SAR ontology and the Wordsmyth dictionary-thesaurus. Evaluation is carried out by comparing the output of the system with key answers of predefined extraction templates. Results on speech transcriptions are comparable to those on written texts in MUC7.

## 1. Introduction

We present a semantic labeling approach for the identification of what we call generalized named entities (GNE) from transcribed conversations. The GNE are selected types of entities same as named entities, however they are not restricted to noun phrases; they can be verbs or adjectives.

The extended semantic tagger is part of a general framework for IE pattern discovery from conversations. Targeted text is transcribed conversational speech which is more complex than transcriptions of Broadcast news (Chincor and al., 1998).In particular, the structural complexity of transcribed conversations such as turn-takings make relevant information scattered through several utterances. Speech disfluencies such as repairs and omissions alter utterances structure and increase the number of ways a given relation may be expressed. Hence, collecting relatively complete set of IE patterns from speech corpora become an even more difficult task than from texts.

Our purpose is to learn IE patterns based on GNE. In particular, we focus on the identification of generalized named entities. The extended semantic tagger is based on a statistical model which combines similarity scores and topical information. Similarity scores help identifying word groups likely to convey information related to the domain, whereas topics help distinguishing GNE from word groups which are of no particular interest.

In section 2., we present the issue of IE pattern discovery for transcribed conversations. Our approach is described in section 3. and the extended semantic tagger and its components are described in section 4. The case study in section 5. shows the results of generalized named entity extraction from transcriptions of telephone conversations in the particular domain of maritime Search and Rescue (SAR). We conclude with some proposals for further improvements.

## 2. IE from transcribed speech

IE is about seeking instances of class of events and relations and extracting their arguments. Despite the maturity of the information extraction (IE) tasks for written texts, IE from transcribed speech is currently restricted to the named entity task (Chincor and al., 1998). IE systems developed for well written texts use patterns based on "subject-verb-object" relation that match the sentence structure. However, whereas this is possible for well written texts where relevant event classes are expressed in a relatively easily recognizable grammatical forms, this is not the case for spontaneous speech. Two necessary hypothesis for syntax driven learning approaches are violated when processing spontaneous conversations: grammatically and locality of informations.

IE from transcribed conversational speech is a two-dimensional problem. The syntactic dimension involves the problem of disfluencies. Edited words, omissions and interruptions are examples of disfluencies that alter the utterance structure causing a significant decrease of performance in part-of-speech tagging and parsing (Charniak and Johnson, 2001). Furthermore, altering the syntactic structure of utterances make syntactic driven learning of extraction patterns difficult if not impossible.

The pragmatic dimension deals with the fact that speech and particularly conversational speech is a highly contextualized activity. Turn-takings, interruptions and overlappings, for example, result in the scattering of relevant information across a series of utterances. Tasks that require shallow or deep understanding of utterances, such as IE, must take into account a larger context than individual utterances.

## 3. IE pattern discovery approach

There has been considerable work on the supervised learning and *quasi* unsupervised learning of IE patterns. Supervised learning approaches use corpora which have been manually annotated to indicate the information to be extracted (Califf and Mooney, 1999; Soderland, 1999).

Quasi unsupervised approaches rely heavily on syntactic information such as "subject-verb-object" relations and on a minimum annotated data; usually named entities to bootstrap the learning process (Riloff, 1998; Yangarber and al., 2000).

As far as we know, very little concern has been given to IE patterns discovery from speech corpora actually limited to Broadcast news. Most of the work has been done on texts and was first introduced to address the problem of portability of IE systems to different application domains. The reasons for this limitation are related to the structural complexity of speech. Two problems arise when learning IE patterns from transcribed conversations. Relevant information can be conveyed through successive turn-takings resulting in scattered informations and disfluencies introduce noise in data. Accordingly, patterns are not observed on utterances but on larger contexts which ensure the completeness and coherence of the conveyed informations; in this case the context is a topic segment.

The approach we present is based on supervised learning from automatically annotated transcribed conversations. Basically, we annotate GNE with domain-specific semantic labels and learn predicate-argument relations that describe IE patterns.

The IE pattern discovery process is divided into four steps. The first one is a pre-analysis of the transcribed conversations. It includes shallow parsing to detect noun groups, verbs and adjectives. The second stage is the topic segmentation and labeling. Topic segments are used as extraction units because they are larger contexts that should ensure complete "predicate-arguments" relations. The topic label represent the word context and is used to distinguish relevant entities from words of no particular interest. The third stage is the extraction of GNE. It includes a process for the recognition of known GNE and another one for the Out-Of-Vocabulary (OOV) GNE. The last stage, the IE pattern learning, is a markov model which takes as input GNE recognized in each topic segment. Figure 3. shows the different modules needed for the IE pattern learning process. In this paper, we only present the third stage which is the extraction of GNE. We tackle the problem of semantic labeling of OOV GNE (section 4.). The IE pattern learning module is left for future work and the others components are described in this section.

### 3.1. Domain knowledge

In IE, domain knowledge has generally been encoded in gazetteers for the named entity extraction task or in ontologies to allow inferences to generate more complex facts. In our approach, we encoded the domain knowledge in an ontology for two reasons:

- Ontologies define explicit hierarchical relations such as IS-A or PART-OF relations that can be used to generalize word classes and reduce their number to enhance the IE pattern learning process.

- They provide an interpretation or grounding of word senses, so that word sense disambiguation problem can be reduced.

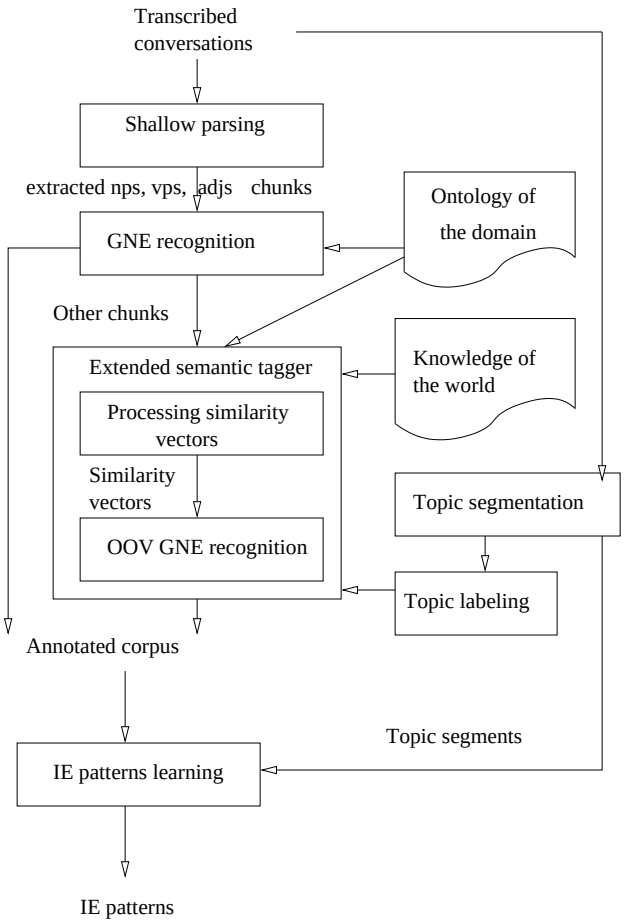


Figure 1: Stages of the IE pattern discovery approach

### 3.2. Knowledge of the world

Dictionaries or lexicons such as WordNet are used to bridge the gap between entities from the corpus which are not described in the ontology of the domain and known entities. Thesaurus in combination with similarity measures have been previously used to enrich ontologies (Stevenson, 2002). In our approach, we used the dictionary-thesaurus Wordsmyth[1] and a similarity measure based on the overlap coefficient to assess the closeness of a word to the domain vocabulary. Figure 2 is an example of a Wordsmyth entry for the word "wonder".

### 3.3. Shallow parsing

Candidates to be tagged are noun groups *np*, verbs *vp* an adjectives *adj*. For this purpose, we used the Brill transformational tagger (Brill, 1992) and the CASS partial parser of Steven Abney (Abney, 1994) to parse the conversations. However, because of the disfluencies encountered in the conversations, many errors occurred when parsing large constructions. So, we reduced the set of grammatical rules used by CASS to cover only minimal chunks and discard large constructions such as VP → H=VX O=NOM? ADV* or noun phrases NP → NP CONJ NP.

[1]URL http://www.wordsmyth.net/.

```
ENT:     wonder
SYL:     won-der
PRO:     wuhn dEr
POS:     intransitive verb
INF:     wondered, wondering, wonders
DEF:     1. to experience a sensa-
tion of admiration or amazement (of-
ten fol. by at):
EXA:     She wondered at his brav-
ery in combat.
SYN:     marvel
SIM:     gape, stare, gawk
DEF:     2. to be curious or skepti-
cal about something:
EXA:     I wonder about his truthfulness.
SYN:     speculate (1)
SIM:     deliberate, ponder, think, re-
flect, puzzle, conjecture
...
```

Figure 2: Description of the dictionnaire-thesaurus Wordsmyth entry for the verb "wonder". This verb express a request for equipement and is tagged as an instance fo the concept STATUS (8-0 Figure 4). Acronymes ENT, SYL, PRO, POS, INF, DEF, EXA, SYN, SIM refers to the entry, syllable, pronunciation, part of speech, flexions, textual definitions, example, synonyms an similar words.

### 3.4. Generalized named entity recognition

This task, like the named entity extraction task, annotates words that are instances of the ontology. Basically, for every chunk, we look for the first match with a concept instance. The match is based on the word and its part-of-speech. When a match succeeds, the semantic tag assigned is the concept of the instance matched. Then, the semantic tag of the head is propagated to the whole chunk as shown in Figure 3.

Matching step 1: ... SN: black thicker $\underbrace{\textit{fog}}_{\text{WEATHER-TYPE}}$ ...
← Propagation
Propagation step: ... SN: $\underbrace{\textit{black thicker fog}}_{\text{WEATHER-TYPE}}$ ...

Figure 3: Output of the named concept extraction process. The semantic tag of the head "fog" is propagated to the whole chunk

In (Boufaden, 2003), we show that the described approach achieves a recall score of 85,3% and a precision score of 94,8%.

### 3.5. Topic segmentation and labeling

The extraction unit we used is the topic segment which is composed of consecutive utterances conveying, in general, at most one piece of information that could be used to fill in a template slot. For this purpose, we developed a topic segmentation system based on a multi-knowledge source modeled by a hidden Markov model. (Boufaden and al., 2001) showed that by using linguistic features modeled by a Hidden Markov Model, it is possible to detect about 67% of topics boundaries.

The topic labeling system has not yet been developed but we are planning to develop it to fully automatically generate word context as described in our approach.

## 4. Extended semantic tagging

Our approach is based on psycholinguistic evidence. It has been shown that, when communicating intentions, speakers select words carefully in order to make the intention recognizable (Levelt, 1993). So, if we consider topics as indicators of communicative intentions, we can assume that given a relevant topic, words with high similarity scores are likely to convey relevant information. Hence, the relevance of a word in a specific domain can be translated into a function of word similarity to domain ontology concepts and the concepts frequency given the topic where the word appears.

In practical terms, the extended semantic tagger is a normalized product of two experts. The first expert is a similarity based model $P(C_t = k|w_t)$ that generates similarity probabilities of concepts to words from similarity scores. Whereas, the second expert is a topic based model $P(C_t = k|T_t)$ that generates concept probabilities given a topic. The product of experts is given by the equation 1.

$$P(C_t = k|w_t, T_t) = \frac{P(C_t = k|w_t)^{\beta_1} P(C_t = k|T_t)^{\beta_2}}{\Sigma_{l=1}^{K} P(C_t = l|w_t)^{\beta_1} P(C_t = l|T_t)^{\beta_2}} \quad (1)$$

$$and$$

$$C^* = \underset{C_t}{\operatorname{argmax}}\, P(C_t|w_t, T_t), P(C^*|w_t, T_t) > \delta \quad (2)$$

$k$ is one from the $K$ concepts of the domain ontology or an Out Of Vocabulary concept (OOV), $P(C_t = k|w_t)$ is the probability that concept $k$ is observed given the word $w_t$ and $P(C_t = k|T_t)$ is the probability that concept $k$ is observed given a topic $T_t$.$\beta_1$ and $\beta_2$ are parameters of the model

Since we are looking for GNE rather than doing only semantic tagging, we empirically determine a threshold to distinguish word groups representing GNE from non relevant words as shown in equation 2.

### 4.1. The similarity based model

The similarity based model generates a vector of similarity scores for each word. It uses a domain ontology and the Wordsmyth dictionary-thesaurus to determine the similarity score between a word and every concept of the domain ontology. They are computed using textual definitions of words as described in Lesk's approach (Lesk, 1996). Technical details of the algorithm used to generate similarity score vectors are described in (Boufaden, 2003). Basically, the similarity score is based on the overlap coefficient similarity measure (Manning and Schutze, 2001). It counts the number of lemmatized content words in common between the textual definition of the word and the concept. In these experiments, we do not address the word sense disambiguation problem and each similarity score is replaced

by the mean of similarity scores of every word sense. We also assume conditional independence between a word and a concept $P(C_k|w(l), w) = P(C_k|w(l))$ where $w(l)$ is a word sense of $w$. In addition, we assume that word senses $w(l)$ are equally probable given a word $w$ (Equation 3).

$$P(w(l)|w) = \frac{1}{|S(w)|} \quad (3)$$

Where $S(w)$ contains the different word senses of $w$ provided by the Wordsmyth dictionary-thesaurus

Hence, $P(C_k|w)$ is given by:

$$P(C_k|w) = \Sigma_{w(l)\in S(w)} P(C_k|w(l))P(w(l)|w) \quad (4)$$

Where $P(C_k|w(l))$ is calculated from similarity scores between the concept $C_k$ given a word sense $w(l)$ of $w$ and $P(w(l)|w)$ is the relative frequency of the word sense $w(l)$ given $w$ from Wordsmyth.

To process $P(C_k|w)$ we added an Out Of Vocabulary (OOV) concept for all the words that have null similarity scores for all SAR concepts. The probabilities are then generated from similarity scores by using a discounting method (Manning and Schutze, 2001).

### 4.2. The topic based model

The topic based model identifies the distribution of concepts given specific topics related to the domain. Basically, for every *event template* (MUC, 1998) we define a topic label. Then, each conversation is divided manually into topic segments and each topic segment is manually labeled with one of the defined topic labels or with the label *other-topic*. Concepts are classified according to equation 5.

$$P(C_t|T_t) = \alpha P_0(C_t) + (1-\alpha)P_1(C_t|T_t) \quad (5)$$

$C_T$ are the ontology concepts, $T_T$ topics. $\alpha$ is the smoothing parameter. $P_0(C_t)$ is the relative frequency and $P_1(C_t|T_t)$ is the relative frequency given a topic.

## 5. Case study: IE from manually transcribed SAR conversations

Our aim is to implement an information extraction system in the domain of Search And Rescue (SAR) from transcribed conversations. The conversations are mostly informative dialogs, where two speakers (a caller `C` and an operator `O`) discuss the conditions and circumstances related to a SAR mission. The conversations are either (1) incident reports, such as reporting missing airplanes or overdue boats, (2) SAR mission plans, such as requesting a SAR airplane or coast guard ships for a mission, (3) debriefings, in which case the results of the SAR mission are communicated. Figure 4 is an excerpt of such conversations. We can see that parts of some utterances were replaced by the word "INAUDIBLE" to indicate segments that have not been transcribed. In the overall corpus, such segments are found in 10% of the utterances. Besides, more than half of the corpus utterances have disfluencies such as repetitions (Ha, do, is there, is there ...) , omissions and interruptions (we've been, _ actually had a ...). There are about 3% transcription errors (such as `flowing` instead of `blowing` in 21-O Figure 4) which mostly occur with relevant words.

The words shown over braces in Figure 4 are the GNE to be extracted. These are, for example, the incident, its location, SAR resources needed for the mission and weather conditions. We can see the role of the topic in distinguishing entities from non relevant words. For example, in utterance 7-C the word "land" is an entity that refers to the STATUS of an airplane[2] having trouble, whereas in utterance 42-O it is of no particular interest.

### 5.1. SAR ontology

We built a SAR ontology using manuals provided by the National Search and Rescue Secretariat (SAR Manual, 2000) and from a sampling of 10 conversations. The ontology is composed of a sampling of key answers of predefined IE template fields such as "radar search", "diving" for means of detection, "drifting", "overdue" for incidents and "wind", "rain", "fog" for weather conditions. All were grouped into 24 semantic classes and organized in IS-A and PART-OF hierarchies. The overall ontology has a maximal depth of three. Each class represents a SAR concept and they are all used to classify entities. For each instance from the ontology we associated a list of synonyms and similar words along with their textual definitions, all extracted from Wordsmyth. Synonyms and similar words were added to increase the effectiveness of the similarity measure used.

### 5.2. Experiments and Results

Experiments were conducted on 4570 words that were manually annotated with SAR concepts and topic labels. 25.3% of these words are GNE. The training corpus represents 65% of the 64 manually transcribed conversations. Relevant topic segments[3] have an average length of 3 utterances. Evaluation is carried out by comparing the output of the system with key answers of predefined extraction templates. A threshold $\delta = 0.35$ was determined empirically. It means that only words that have $P(C_t|w_t, T_t) > 0.35$ are considered as GNE. Table 1 shows the precision and recall of the extended semantic tagger and the similarity based model. For the topic based model we proceed to the evaluation of the classification error. All the modules were tested on manually segmented conversations.

The major result is an assessement of the feasibility of the GNE extraction task. Our system achieves an F-score[4] of 86% which is not as good as F-scores of NEE from transcribed speech around 93% (Miller and al., 1999). However, Broadcast News are well written texts read by a speaker and can not be considered as spontaneous speech. On the other hand, our texts are spontaneous conversations with disfluencies that significantly decrease the part-of-speech tagging performance which results in increasing the semantic labeling error. Besides, since named entities are a subset of the generalized named entities we consider

[2]The airplane is actually considered as a missing object

[3]Relevant topics are topic segments that are not labeled with the 'other-topic' tag.

[4]F-score used is $F = \frac{(\beta+1)P.R}{\beta^2.P+R}$ and $\beta = 0.5$

...
———————— INCIDENT————————-

7-C: On the way to go, he had to **land** in **emergency** in the **South East Coast of Newfoundland**.
STATUS INCIDENT LOCATION

...
———————— SEARCH UNIT————————-

12-O: They did **a radar search** for us in **the 3 other surfaces**.
TASK LOCATION

13-C: Hum, hum.
———————— SEARCH UNIT————————-

18-O: And I am **wondering** about the possibility of **outputting** an **Aurora** in there for **radar search**.
STATUS STATUS SAR-AIRCRAFT TASK

...
———————— MISSION————————-

21-O: They got **a South East** to be **flowing** there and it's just **gonna** be **black thicker fog** the whole, **whole South Coast**.
DIRECTION STATUS STATUS WEATHER LOCATION

22-C:OK.
———————— OTHER-TOPIC————————-

42-O: Now, the question he had was is there some place for a small helicopter to land there,
if he was to get something else or somebody else to take him in there ?

...
———————— SEARCH UNIT————————-

56-C: Ha, they **should go** **to get going** at **first light**.
STATUS STATUS TIME

...

Figure 4: An excerpt of a conversation reporting an emergency landing and a request for an SAR airplane (Aurora). Numbers are utterances position in the conversation. The words in bold are extracted GNE. The tag below each bold chunk is an SAR concept from the ontology which we want to identify. Lines are boundaries of topics which were added manually (MISSION, INCIDENT, SEARCH UNIT, OTHER-TOPIC)

T1:INCIDENT

| | | |
|---|---|---|
| *1* | *Initial alert* | `emergency landing` |
| *3* | *Location* | `South East Coast of New-foundland` |
| *4* | *Date* | |
| *5* | *Missing object* | `airplane` |
| *7* | *Weather conditions* | `WEATHER1` |

T2:WEATHER CONDITIONS

| | | |
|---|---|---|
| *1* | *Id* | `WEATHER1` |
| *2* | *Condition* | `black thicker fog` |
| *3* | *Wind direction* | `South East` |
| *4* | *Wind speed* | |
| *5* | *Visibility* | |

Figure 5: Two filled templates from the conversation in Figure 4: the event template "INCIDENT" and the object template "WEATHER CONDITIONS".

our results comparable to those on NEE from Broadcast News.

| | $P(C_t\|T_t)$ | $P(C_t\|w_t)$ | $P(C_t\|T_t, w_t)$ |
|---|---|---|---|
| Precision | 48.8% | 61.0% | 76.8% |
| Recall | | 55.2% | 56.7% |

Table 1: $P(C_t|T_t)$ is topic-based model, $P(C_t|w_t)$ the similarity-based model and $P(C_t|T_t, w_t)$ the combined model. Results of the combined model are obtained with a threshold of $\delta = 0.35$. To compare the similarity based model with the combined model we tested word groups with $P(C_t|w_t) > 0.005$.

As well, results show the effectiveness of the combined model over the similarity based model $P(C_t|w_t)$. Despite the poor performance of the topic based model $P(C_t|T_t)$, it improves the detection of OOV GNE by $25.9\%$.

## 6. Conclusion

Named entity extraction (NEE) is an important stage for text based IE systems because it's a relatively easy task that has proved to be helpful for IE pattern learning. However, because of the structural complexity of transcribed speech, we believe that moving beyond named entities to identify GNE would be more helpful for the IE pattern discovery task applied to conversations.

In this paper, we experiment on the recognition of GNE related to a particular domain. The extended semantic tagger used is a stochastic model which combine similarity scores and topical information to generate semantic labels drawn from a domain ontology we designed. It is part of an ongoing work that aim to develop an IE pattern discovery method that learns predicate-arguments relations from a corpus annotated with domain-specific semantic tags.

Results of the experiments are not as good as those of related works in named entities extraction or on shallow semantic parsing (Gildea and Jurafsky, 2002). However, we believe that IE pattern based on domain-specific semantic tags is a way to get around the structural complexity of conversations.

The system being at a preliminary stage, there is room for further improvements including better smoothing in the generation of $P(C_k|w_t)$ from similarity scores. For the case study, we have worked on manually segmented conversations with manually annotated topic labels. But, we are planning to develop a system to automatically label topic segments as generated by the system described in (Boufaden and al., 2001). The last step in our project is to learn a set of IE patterns to validate our approach.

## 7. References


S. Abney. 1994. Partial parsing. Tutorial given at ANLP.

E. Brill. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

MUC 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Morgan Kauffman.

Boufaden, N., G. Lapalme, and Y. Bengio, 2001. Topic Segmentation : A First Stage to Dialog-based Information Extraction. In *Natural Language Processing Pacific Rim Symposium, NLPRS'01*.

Boufaden, N. An ontology-based semantic tagger for ie system. In *ACL Student Workshop*, pages 7–14, Sapporo, Japon, Juillet 2003.

Chincor, N., Robinson P., and Brown E., 1998. HUB-4 Named Entity Task Definition Version 4.8. Technical report.

Califf, E. M. and Mooney R., 1999. Relational Learning of Pattern Match Rules for Information Extraction. In *16th National Conference on Artificial Intelligence AAA-I 99*.

Charniak, E. and Johnson M., 2001. Edit Detection and Parsing for Transcribed Speech. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*.

Gildea, D. and Jurafsky D., 2002. *Automatic Labeling of Semantic Roles*, volume 28(3) of *Computational Linguistics*. pages 245–288.

Grishman, R., 1997. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, SCIE-97*, volume 1299 of *Lecture Notes on Computer Sience*, chapter two. Springer Verlag, pages 10–27.

Lesk, M., 1996. Automatic Sense Disambiguation:How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the SIGDOC Conference*.

Stevenson, M., 2002. Combining Disambiguation Techniques to Enrich an Ontology. In *15th Conference on Artificial Intelligence EACI-02 workshop on "Machine Learning and natural Processing for Ontology Engineering*. lyon, France.

Manning, C. D. and Schütze, H., 2001. *Foundations of Statistical Natural Language Processing*, chapter Word Sense Disambiguiation. The MIT Press Cambridge, Massachussets London England, pages 294–303.

Miller, D., Schwartz D.R., Weischedel, R., and Stone, R.. Named entity extraction from broadcast news. In *Proceedings of DARPA Broadcast News Workshop*.

Search And Rescue Manual, 2000. Fisheries and Oceans Canada, Canadian Coast Guard, Search and Rescue, 2000. *SAR Seamanship Reference Manual*, Canadian Government Publishing, Public Works and Government Services Canada edition, November. ISBN 0-660-18352-8.

Yangarber, R., Grishman R., Tapanainen P., and Hutunen S., 2000. Automatic Acquisition od Domain Knowledge for Information Extraction. In *18th COLING Conference*. Germany.

Riloff, E., 1998. Automatically generating extraction patterns from untagged tex. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*.

Soderland, S. , 1999. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34:233–272.

Levelt, W.J.M., 1993. *Speaking: From Intention to Articulation*. MIT Press.