An Ontology-based Semantic Tagger for IE system

Narjès Boufaden
Department of Computer Science
Université de Montréal
Quebec, H3C 3J7 Canada
boufaden@iro.umontreal.ca

Abstract

In this paper, we present a method for the semantic tagging of word chunks extracted from a written transcription of 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 annotate parts of texts with concepts from a SAR ontology. Our approach combines two knowledge sources a SAR ontology and the Wordsmyth dictionary-thesaurus, and it uses a similarity measure for the classification. Evaluation is carried out by comparing the output of the system with key answers of predefined extraction templates.

1 Introduction

This work is a part of a project aiming to implement an information extraction (IE) system in the field of maritime Search And Rescue (SAR). It was originally conducted by the Defense Research Establishment Valcartier (DREV) to develop a decision support tool to help in producing SAR plans given the information extracted by the SAR IE system from a collection of transcribed dialogs. The goal of our project is to develop a robust approach to extract relevant words for small-scale corpora and transcribed speech dialogs. To achieve this task, we developed a semantic tagger which annotates words with domain-specific informations and a selection process to extract or reject a word according to the semantic tag and the context. The rationale behind our approach, is that the relevance of a word depends strongly on how close it is to the SAR domain and its context of use. We believe that reasoning on semantic tags instead of the word is a way of getting around some of the problems of small-scale corpora.

In this paper, we focus on semantic tagging based on a domain-specific ontology, a dictionary-thesaurus and the overlapping coefficient similarity measure (Manning and Schutze, 2001) to semantically annotate words.

We first describe the corpus (section 2), then the overall IE system (section 3). Next we explain the different components of the semantic tagger (section 4) and we present the preliminary results of our experiments (section 5). Finally we give some directions for future work (section 6).

2 Corpus

The corpus is a collection of 95 manually transcribed telephone conversations (about 39,000 words). They 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 persons or overdue boats, (2) SAR mission plans, such as requesting an SAR airplane or coast guard ships for a mission, or (3) debriefings, in which case the results of the SAR mission are communicated. They can also be a combination of the three kinds. Figure 1 is an excerpt of such conversations. We can notice many disfluencies

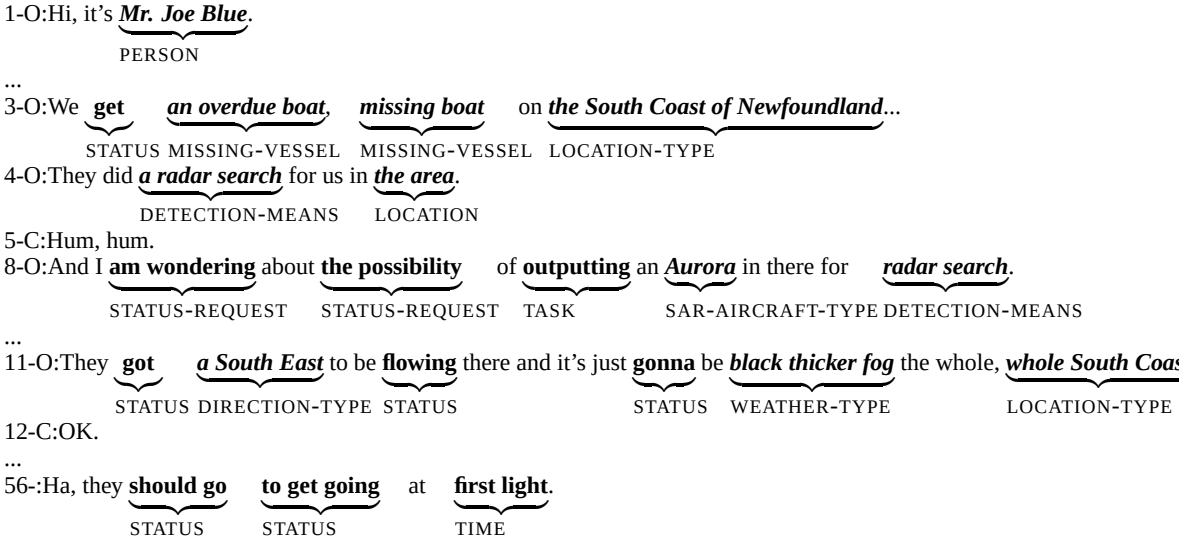

Figure 1: An Excerpt of a conversation reporting an overdue vessel:the incident, a request for an SAR airplane (Aurora) and the use of another SAR airplane (king Air). The words in bold are candidates for the extraction. The tag below each bold chunk is a domain-specific information automatically generated by the semantic tagger. Chunks like **possibility**, **go**, **flowing** and **first light** are annotated by using sense tagging outputs. Whereas chunk such as ***Mr. Joe Blue***, the ***South coast of Newfoundland*** and ***Aurora*** are annotated by the named concept extraction process.

(Shriberg, 1994) such as repetitions (`13-O: Ha, do, is there, is there ...`), omissions and interruptions (`3-O: we've been, actually had a ...`). And, there is about 3% of transcription errors such as `flowing` instead of `blowing` (11-O Figure 1).

The underlined words are the relevant informations that will be extracted to fill in the IE templates. They are, for example, the incident, its location, SAR resources needed for the mission, the result of the SAR mission and weather conditions.

3 Overall system

The information extraction system is a four stage process (Figure 2). It begins with the extraction of words that could be candidates to the extraction (stage I). Then, the semantic tagger annotates the extracted words (stage II). Next, given the context and the semantic tag a word is extracted or rejected (stage III). Finally, the extracted words are used for the coreference resolution and to fill in IE templates (stage IV). The knowledge sources used for the IE task are the SAR ontology and the Wordsmyth dictionary-thesaurus[1].

In this section we describe the extraction of candidates, the SAR ontology design and the topic segmentation which have already been implemented. We leave the description of the topic labeling, the selection of relevant words and the template generation to future work. The semantic tagger, is detailed in section 4.

3.1 Extraction of candidates

Candidates considered in the semantic tagging process are noun phrases NP, proposition phrases PP, verb phrases VP, adjectives ADJ and adverbs ADV. To gather these candidates we used the Brill transformational tagger (Brill, 1992) for the part-of-speech step and the CASS partial parser for the parsing step (Abney, 1994). However, because of the disfluencies (repairs, substitutions and omissions) 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 $\rightarrow$ VX NP? ADV* or noun

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

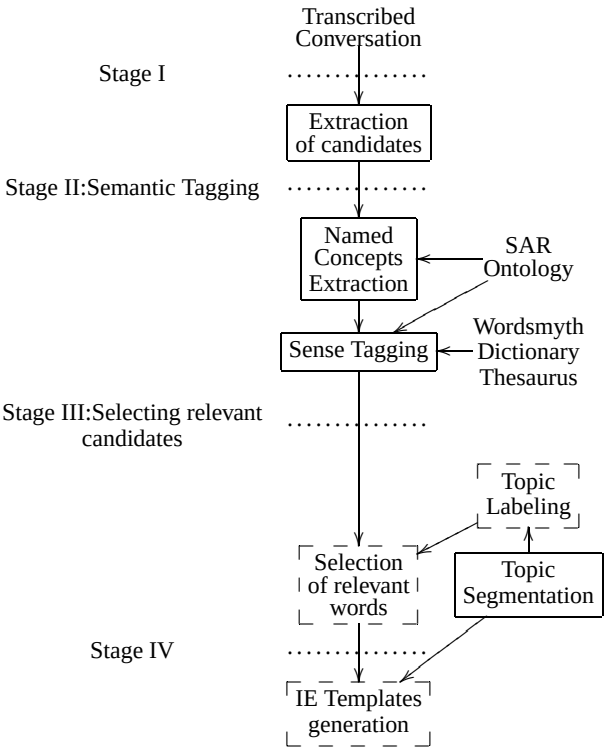

Figure 2: Main stages of the full SAR information extraction system. Dashed squares represent processes which are not developed in this paper.

phrases NP $\rightarrow$ NP CONJ NP. The evaluation of the semantic tagging process shows that about 14.4% of the semantic annotation errors are partially due to part-of-speech and parsing errors.

3.2 Topic segmentation

Topic segmentation takes part to several stages in our IE system (Figure 2). Dialogue-based IE systems have to deal with scattered information and disfluencies. Question-answer pairs, widely used in dialogues, are examples where information is conveyed through consecutive utterances. By dividing the dialog into topical segment, we want to ensure the extraction of coherent and complete key answers. Besides, topic segmentation is a valuable preprocessing for coreference resolution, which is a difficult task in IE. Hence, for the extraction of relevant candidates and the coreference resolution which is part of the template generation stage (Figure 2), we use topic segment as context instead of the utterance or a word window of arbitrary size.

The topic segmentation system we developed is based on a multi-knowledge source modeled by a hidden Markov model. (N. 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.

3.3 The SAR ontology

The SAR ontology is an important component of our IE system. We build it using domain related informations such as airplane names, locations, organizations, detection means (`radar search`, `diving`), status of a SAR mission (`completed`, `continuing`, `planned`), instance of maritime incidents (`drifting`, `overdue`) and weather conditions (`wind`, `rain`, `fog`). All these informations were gathered from SAR manuals provided by the National Search and Rescue Secretariat (SARManual, 2000) and from a sample of conversations (10 conversations about 10% of the corpus) to enumerate the different status informations.

Our ontology was designed for two tasks of the semantic tagging:

1. Annotate with the corresponding concept all the extracted words that are instances of the ontology. This task is achieved by the named concept extraction process (section 4.1).

2. For each word not in the ontology, generate a concept-based representation composed of similarity scores that provide information about the closeness of the word to the SAR domain. This is achieved by the sense tagging process (section 4.2).

In addition to SAR manuals and corpus, we used the IE templates given by the DREV for the design of the ontology. We used a combination of the top-down and bottom-up design approaches (Fridman and Hafner, 1997). For the former, we used the templates to enumerate the questions to be covered by the ontology and distinguish the major top level classes (Figure 4). For the latter, we collected the named entities along with airplane names, vessel types, detection means, alert types and incidents. The taxonomy is based on two hierarchical relations: the *is-a* relation and the *part-of* relation. The *is-a* relation is used for the semantic tagging. Whereas, the

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

Figure 3: A fragment of the Wordsmyth dictionary-thesaurus entry of the verb **wonder** which is a verb describing a STATUS-REQUEST concept (8-0 Figure 1). The ENT, SYL, PRO, POS, INF, DEF, EXA, SYN, SIM acronyms are respectively the entry, the syllable, the pronunciation, the part-of-speech, inflexion form, textual definition, example, synonim words and similar words fields. To build the SAR ontology we used the information given in the fields DEF, SYN and SIM. Whereas, to compute the similarity scores we used only the information of the DEF field.

part-of relation will be used in the template generation process.

The overall ontology is composed of 31 concepts. In the *is-a* hierarchy, each concept is represented by a set of instances and their textual definitions. For each instance we added a set of synonyms and similar words and their textual definitions to increase the size of the SAR vocabulary which was found to be insufficient to make the sense tagging approach effective.

All the synonyms and similar words along with their definitions are provided by the Wordsmyth dictionary-thesaurus. Figure 3 is an example of Wordsmyth entries. Only textual definitions that fit the SAR context were kept. This procedure increases the ontology size from 480 for a total of 783 instances.

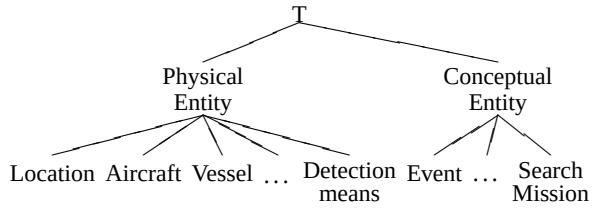

Figure 4: Fragment of the *is-a* hierarchy. Location, Aircraft ... are concepts of the ontology

4 Semantic tagging

The purpose of the semantic tagging process is to annotate words with domain-specific informations. In our case, domain-specific informations are the concepts of the SAR ontology. We want to determine the concept C_k which is semantically the most appropriate to annotate a word w. Hence, we look for C^* which has the highest similarity score for the word w as shown in equation 1.

$$C^* = \underset{C_k}{\operatorname{argmax}}\, sim(w, C_k) \quad (1)$$

Basically, our approach is a two part process (figure 2). The named concept extraction is similar to named entity extraction based on gazetteer (MUC, 1991). However it is a more general task since it also recognizes entities such as, aircraft names, boat names and detection means. It uses a finite state automaton and the SAR ontology to recognize the named concepts.

The sense tagging process generates a based-concept representation for each word which couldn't be tagged by the named concept extraction process. The concept-based representation is a vector of similarity scores that measures how close is a word to the SAR domain. As we mentioned before (section 1), the concept-based representation using similarity

scores is a way to get around the problem of small-scale corpora. Because we assume that the closer a word is to an SAR concept, the more relevant it is, this process is a key element for the selection of relevant words (figure 2). In the next two sections, we detail each component of the semantic tagger.

4.1 Named concept extraction

This task, like the named entity extraction task, annotates words that are not instances of the ontology. Basically, for every chunk, we look for the first match with an instance concept. 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. The propagation of the semantic tag is done by a two level automaton. The first level propagates the semantic tag of the head to the whole chunk. The second level deals with cases where the first level automaton fails to recognize collocations which are instances of the ontology.

These cases occur when :

- the syntactic parser fails to produce a correct parse. This mainly happens when the part of speech tag isn't correct because of disfluencies encountered in the utterance or because of transcription errors.
- the grammatical coverage is insufficient to parse large constructions.

Whenever one of these reasons occur, the second level automaton tries to match chunk collocations instead of individual chunks. For example, the chunk `Rescue Coordination Centre` which is an organization, is an example where the parser produces two NP chunks (NP1:`Rescue Coordination` and NP2:`Centre`) instead of only one chunk. In this case, the first level automaton fails to recognize the organization. However, in the second level automaton, the collocation NP1 NP2 is considered for matching with an instance of the concept *organization*. Figure 5 shows two output examples of the named concept extraction.

Finally, if the automaton fails to tag a chunk, it assigns the tag OTHER if it's an NP, OTHER-PROPERTIES if it's a ADJ or ADV and OTHER-STATUS if it's a VP.

4.2 Sense tagging

Sense tagging takes place when a chunk is not an instance of the ontology. In this case, the semantic tagger looks for the most appropriate concept to annotate the chunk (equation 1). However, a first step before annotation is to determine what word sense is intended in conversations. Many studies (Resnik, 1999; Lesk, 1986; Stevenson, 2002) tackle the sense tagging problem with approaches based on similarity measures. Sense tagging is concerned with the selection of the right word sense over all the possible word senses given some context or a particular domain. Our assumption is that when conversations are domain-specific, relevant words are too. It means that sense tagging comes back to the problem of selecting the closer word sense with regard to the SAR ontology. This assumption is translated in equation 2.

$$w^* = \underset{w(l)}{\operatorname{argmax}} \frac{1}{N_l} \Sigma_{\text{all concepts k}} sim(w(l), k) \tag{2}$$

Where N_l is the number of positive similarity scores of the $w(l)$ similarity vector. $w(l)$ is the word w given the word sense l. The closer word sense w^* is the highest mean computed from element of the $w(l)$ similarity vector.

In what follows, we explain how are generated the similarity vectors and the result of our experiments.

4.3 Similarity vector representation

A similarity vector is a vector where each element is a similarity score between a $word(l)$ (the word w given the sense word l) and a concept C_k from the SAR ontology. The similarity score is based on the overlap coefficient similarity measure (Manning and Schutze, 2001). This measure counts the number of lemmatized content words in common between the textual definition of the word and the concept. It is defined as :

$$sim(w(l), C_k) = \frac{\mid D_{w(l)} \mid \cap \mid D_{C_k} \mid}{min(\mid D_{w(l)} \mid, \mid D_{C_k} \mid)} \tag{3}$$

where $D_{w(l)}$ and D_{C_k} are the sets of lemmatized content words extracted from the textual definitions

```
3-O:an overdue boat
VESSEL:[dt,an],[OTHER-PROPERTIES,overdue],[VESSEL,boat]

11-O:black thicker fog
WEATHER-TYPE:[COLOR-TYPE,black],[OTHER-PROPERTIES,thicker],[WEATHER-TYPE,fog]
```

Figure 5: Output of the named concept extraction process. For both chunks the head semantic tag is propagated to the whole chunk

for each concept C_k of the SAR ontology; $C_k \in \{$*incident,detection-means,status*$\dots\}$
  for each instance I_j of C_k; $I_j \in \{$*broken,missing,overdue*$\dots\}$ for the concept *incident*
    for each synonym S_i of I_j; $S_i \in \{$*smach,crack*$\dots\}$ for the instance *broken*
      $sim(w(l), S_i)$=$\frac{|D_{w(l)}| \cap |D_{S_i}|}{min(|D_{w(l)}|, |D_{S_i}|)}$
    end
    $\vec{v_j} \stackrel{\text{def}}{=} (sim(w(l), S_1), \dots, sim(w(l), S_{N_j}))$
    $sim(w(l), I_j)$=$mediane(\vec{v_j})$
  end
  $\vec{v_k} \stackrel{\text{def}}{=} (sim(w(l), I_1), \dots, sim(w(l), I_{M_k}))$
  $sim(w(l), C_k)$=$max(\vec{v_k})$
end
$\vec{v}^{w(l)} \stackrel{\text{def}}{=} (sim(w(l), C_1), \dots, sim(w(l), C_M))$

Figure 6: Similarity measure algorithm. N_j is the number of synonyms for the instance I_j, M_k the number of the instance for the concept C_k and M the number of concepts in the ontology.

of $w(l)$ and C_k. The textual definitions are provided by the Wordsmyth thesaurus-dictionary.

However, since we have represented each concept by a set of instances and their synonyms in the SAR ontology (section 3.3), we modified the similarity measure to take into account the textual definition of concept instances and their synonyms. Basically, we compute the similarity score between $w(l)$ and each synonym S_i of a concept instance I_j. Then, the similarity score between $w(l)$ and the instance concept I_j is the median of the resulting similarity vector representing the similarity scores over all the synonyms. Finally, the similarity score between a concept C_k and $w(l)$ is the highest similarity score over all the concept instances. The algorithm describing these steps is given in Figure 6.

5 Preliminary results and discussion

The evaluation of the semantic tagging process was done on 521 extracted chunks (about 10 conversations). Only relevant chunks where considered for the evaluation. The evaluation criteria is an assessment about the appropriateness of the selected concept to annotate the word. For example, the concept *time* is appropriate for the word *first light*, whereas the concept *incident* is not for the word *detachment* which is closer to the *search_unit* concept.

Chunk	Mean sim	Nearest concepts
get	0.5	0.5 - status
suitable	0.53	0.53 - status
possibility	0.14	0.29-status;0.25-person
first light	0.25	0.25 - time

Table 1: Output samples from the semantic tagger. Mean sim is the mean of the similarity scores. It is the selection criteria used to choose the closest word sense.

Table 2 shows the recall and precision scores for each component and for the overall semantic tagger. The third column shows the input error rates for each component. The error rate in the first row comprises

Process	Recall	Precis.	Inp.Err
Named concept extraction	85.3%	94.8%	7.3%
Semantic tagger using sense tagging output	93.5%	72.6%	11.3%
Average performance of the semantic tagger	89.4%	83.7%	8.3%

Table 2: Precision and Recall scores for each components of the semantic tagger

error rates of the part-of-speech tagger, the parsing and the manual transcription. The error rate in the second row are mostly part-of-speech errors. In spite of the significant error rate, the approach based on partial parsing is effective. The use of a minimal grammar coverage to produce chunks reduced considerably the parsing error rate.

As far as we know, no previous published work on domain-specific WSD for speech transcriptions has been presented, although, word sense disambiguation is an active research field as demonstrated by SENSEVAL competitions[2]. Hence it is difficult to compare our results to similar experiments. However, some comparative studies (Maynard and Ananiadou, 1998; Li Shiuan and Hwee Tou, 1997) on domain-specific well-written texts show results ranging from 51,25% to 73,90%. Given the fact that our corpus is composed of speech transcriptions with the effect of increasing parsing errors, we consider our results to be very encouraging.

Finally, results reported in Table 2 should be regarded as a basis for further improvement. In particular, the selection criteria in the sense tagging process could be improved by considering other measures than the mean of all similarity scores as shown in equation 2.

6 Future work

Extraction of relevant words is a hub for several applications such as question-answering and summarization. It is based on semantically tagging words and selecting the most relevant ones given the context. In this paper, we developed a semantic tagging approach that uses a domain-specific ontology, a dictionary-thesaurus and the overlapping coefficient similarity measure to annotate words. We have shown how the use of concepts to represent words can alleviate the problem of small-scale corpora for the selection of relevant words.

The next step in our project is the selection of relevant words given the concepts annotating them and the topic segments where they appear. Selection will be based on a combination of a probabilistic model taking into account the probability of observing a concept given a word and the probability of observing that concept given a relevant topic.

[2]URL:http://www.senseval.org/.

Acknowledgments

We are grateful to Robert Parks at Wordsmyth organization for giving us the electronic Wordsmyth version. Thanks to the Defense Research Establishment Valcartier for providing us with the dialog transcriptions and to National Search and rescue Secretariat for the valuable SAR manuals.

References

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

N. Boufaden G. Lapalme and Y. Bengio. 2001. Topic segmentation : A first stage to dialog-based information extraction. In *Natural Language Processing Rim Symposium, NLPRS'01*, pages 273–280.

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

Manual. 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.

N. Fridman and C.D. Hafner. 1997. State of the art in ontology design. *AI Magazine*, 18(3):53–74.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC Conference*, pages 24–26, Toronto, Canada.

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

MUC,1991. *Proceedings of the Third Message Understanding Conference*. Morgan Kaufman.

D. Maynard and S. Ananiadou, 1998. 1998. Term Sense Disambiguation using a Domain-Specific Thesaurus. In *Proceedings of 1st International Conference on Language Resource and Evaluation (LREC)*, Granada, Spain.

P. Resnik, 1999. *Natural Language Processing using Very Large Corpora*, chapter Disambiguating Noun Groupings with Respect to WordNet senses, pages 77–98. S. Amstrong, K. Church, P. Isabelle, S. Manzi, E, Tzoukermann and D. Yarowsky, kluwer Academic Press edition.

E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. Thèse de doctorat, University of California at Berkeley.

P. Li Shiuan and N. Hwee Tou 1997. Domain-Specific Semantic Class Disambiguation Using Wordnet. In *Proceedings of the fifth Workshop on Very Large Corpora*, pages 56–64, Beijing and Hong Kong.

M. Stevenson. 2002. Combining Disambiguation Techniques to Enrich an Ontology. In *Proceedings of the Fifteen European Conference on Artificial Intelligence, workshop on Machine Learning and Natural Language Processing for Ontology Engineering*, Lyon,France.