Automatic Construction of Frame Representations for Spontaneous Speech in Unrestricted Domains

Klaus Zechner
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
zechner@cs.cmu.edu

Abstract

This paper presents a system which automatically generates shallow semantic frame structures for conversational speech in unrestricted domains.

We argue that such shallow semantic representations can indeed be generated with a minimum amount of linguistic knowledge engineering and without having to explicitly construct a semantic knowledge base.

The system is designed to be robust to deal with the problems of speech dysfluencies, ungrammaticalities, and imperfect speech recognition.

Initial results on speech transcripts are promising in that correct mappings could be identified in 21% of the clauses of a test set (resp. 44% of this test set where ungrammatical or verb-less clauses were removed).

1 Introduction

In syntactic and semantic analysis of spontaneous speech, little research has been done with regard to dealing with language in unrestricted domains. There are several reasons why so far an in-depth analysis of this type of language data has been considered prohibitively hard:

- inherent properties of spontaneous speech, such as dysfluencies and ungrammaticalities (Lavie, 1996)
- word accuracy being far from perfect (e.g., on a typical corpus such as SWITCHBOARD (SWBD) (Godfrey et al., 1992), current state-of-the-art recognizers have word error rates in the range of 30–40% (Finke et al., 1997))
- if the domain is unrestricted, manual construction of a semantic knowledge base with reasonable coverage is very labor intensive

In this paper we propose to combine methods of partial parsing ("chunking") with the mapping of the verb arguments onto subcategorization frames that can be extracted automatically, in this case, from WordNet (Miller et al., 1993). As preliminary results indicate, this yields a way of generating shallow semantic representations efficiently and with minimal manual effort.

Eventually, these semantic structures can serve as (additional) input to a variety of different tasks in NLP, such as text or dialogue summarization, information gisting, information retrieval, or shallow machine translation.

2 Shallow Semantic Structures

The two main representations we are building on are the following:

- *chunks*: these correspond mostly to basic (i.e., non-attached) phrasal constituents
- *frames*: these are built from the parsed chunks according to subcategorization constraints extracted from the WordNet lexicon

The *chunks* are defined in a similar way as in (Abney, 1996), namely as "non-recursive phrasal units"; they roughly correspond to the standard linguistic notion of constituents, except that there are no attachments made (e.g., a PP to a NP) and that a verbal chunk does *not* include any of its arguments but just consists of the verbal complex (auxiliary/main verb), including possibly inserted adverbs and/or negation particles.

All *frames* are being generated on the basis of "short clauses" which we define as minimal clausal units that contain at least one subject and an inflected verbal form.[1] [2]

To produce the list of all possible subcategorization frames, we first extracted all verbal tokens from the tagged SWITCHBOARD corpus and then retrieved the frames from WordNet. Table 1 provides a summary of this pre-calculation.

[1] This means in effect that relative clauses will get mapped separately. They will, however, have to be "linked" to the phrase they modify.

[2] We are also considering to take even shorter units as basis for the mapping that would, e.g., include non-inflected clausal complements. The most convenient solution has yet to be determined.

Verbal tokens	4428
Different lemmata	2464
Senses in all lemmata	8523
Avg. senses per lemma	3.46
Total number of frames	15467
Avg. frames per sense	1.81

Table 1: WordNet: verbal lemmata, senses, and frames

3 Resources and System Components

We use the following resources to build our system:

- the SWITCHBOARD (SWBD) corpus (Godfrey et al., 1992) for speech data, transcripts, and annotations at various levels (e.g., for segment boundaries or parts of speech)
- the JANUS speech recognizer (Waibel et al., 1996) to provide us with input hypotheses
- a part of speech (POS) tagger, derived from (Brill, 1994), adapted to and retrained for the SWITCHBOARD corpus
- a preprocessing pipe which cleans up speech dysfluencies (e.g., repetitions, hesitations) and contains a segmentation module to split the speech recognizer turns into short clauses
- a chart parser (Ward, 1991) with a POS based grammar to generate the *chunks*[3] (phrasal constituents)
- WordNet 1.5 (Miller et al., 1993) for the extraction of subcategorization (subcat) frames for all senses of a verb (including semantic features, such as "animacy")
- a mapper which tries to find the "best match" between the chunks found within a short clause and the subcat frames for the main verb in that clause

The major blocks of the system architecture are depicted in Figure 1.

We want to stress here that except for the development of the small POS grammar and the frame-mapper, the other components and resources were already present or quite simple to implement. There has also been significant work on (semi-)automatic induction of subcategorization frames (Manning, 1993; Briscoe and Carroll, 1997), such that even

[3]More details about the chunk parser can be found in (Zechner, 1997).

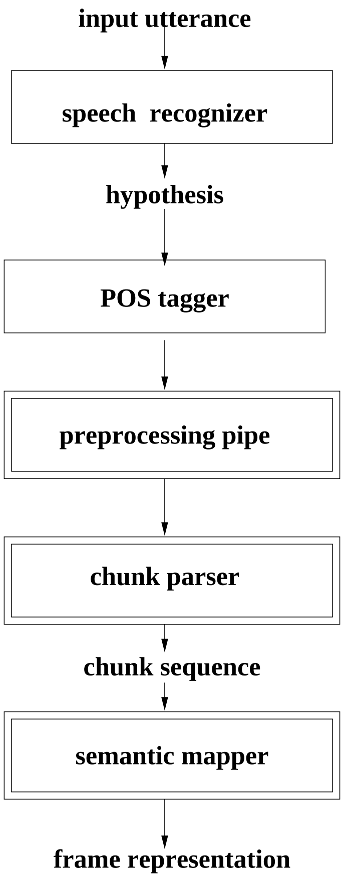

Figure 1: Global system architecture

without the important knowledge source from WordNet, a similar system could be built for other languages as well. Also, the Euro-WordNet project (Vossen et al., 1997) is currently underway in building WordNet resources for other European languages.

4 Preliminary Experiments

We performed some initial experiments using the SWBD transcripts as input to the system. These were POS tagged, preprocessed, segmented into short clauses, parsed in chunks using a POS based grammar, and finally, for each short clause, the frame-mapper matched all potential arguments of the verb against all possible subcategorization frames listed in the lemmata file we had precomputed from WordNet (see section 2).

In total we had over 600000 short clauses, containing approximately 1.7 million chunks. Only 18 different chunk patterns accounted for about half of these short clauses. Table 2 shows these chunk

main verb present?	frequency	chunk sequence
no	83353	— (noises/hesit.)
no	36731	aff
no	33182	conj
yes	29749	np vb
yes	19176	np vb np
no	13834	np
no	13623	conj np
yes	12220	conj np vb
yes	11038	conj np vb np
yes	7649	np vb adjp
yes	7092	np vb pp
yes	5552	np vbneg
no	5044	advp
yes	4926	np vb np pp
no	4079	pp
yes	3999	conj np vb pp
yes	3998	conj np vb adjp
yes	3996	np vb advp

Table 2: Most frequent chunk sequences in short clauses

patterns and their frequencies.[4] Most of these contain main verbs and hence can be sensibly used in a mapping procedure but some of them (e.g., `aff`, `conj`, `advp`) do not. These are typically backchannellings, adverbial comments, and colloquial forms (e.g., "yeah", "and...", "oh really"). They can be easily dealt with a preprocessing module that assigns them to one of these categories and does not send them to the mapper.

Another interesting observation we make here is that within these most common chunk patterns, there is only *one* pattern (`np vb np pp`) which could lead to a potential PP-attachment ambiguity. We conjecture that this is most probably due to the nature of conversational speech which, unlike for written (and more formal) language, does not make too frequent use of complex noun phrases that have one or multiple prepositional phrases attached to them.

We selected 98 short clauses randomly from the output to perform a first error analysis.

The results are summarized in Table 3. In over 21% of the clauses, the mapper finds at least one mapping that is correct. Another 23.5% of the clauses do not contain any chunks that are worth to be mapped in the first place (noises, hesitations), so these could be filtered out and dealt with entirely before the mapping process takes place, as we mentioned earlier. 28.6% of the clauses are in some sense incomplete, mostly they are lacking a main verb which is the crucial element to get the mapping procedure started. We regard these as "hard" residues, including well-known linguistic problems such as ellipsis, in addition to some spoken language ungrammaticalities. The last two categories (26.6% combined) in the table are due to the incompleteness and inaccuracies of the system components themselves.

To illustrate the process of mapping, we shall present an example here, starting from the POS-tagged utterance up to the semantic frame representation:[5] [6]

```
short clause, annotated with POS:
 i/PRP will/AUX talk/VB
 to/PREP you/PRPA again/RB

LEMMA/token (of main verb):
 talk/talk

parsed chunks:
 -np-vb-pp-advp

parsed sequence to map:
 -NP-VBZ-PP

WordNet frames:
:1-INAN-VBZ:1-ANIM-VBZ:1-INAN-IS-VBG-PP
      :1-ANIM-VBZ-PP:1-ANIM-VBZ-TO-ANIM
:2-ANIM-VBZ:2-ANIM-VBZ-PP
:3-ANIM-VBZ:3-ANIM-VBZ-INAN
:4-ANIM-VBZ
:5-ANIM-VBZ
:6-ANIM-VBZ:6-INAN-VBZ-TO-ANIM
      :6-ANIM-VBZ-ON-INAN

Potential mappings (found by mapper):
map. 1: 1-NP-VBZ (1-INAN-VBZ)
map. 2: 1-NP-VBZ (1-ANIM-VBZ)
map. 3: 1-NP-VBZ-PP (1-ANIM-VBZ-PP)
map. 4: 1-NP-VBZ-PP (1-ANIM-VBZ-TO-ANIM)
(...)

Frame representation (for mapping 4):
 [agent/an] (i/PRP)
```

[4] Chunk abbreviations: conj=conjunction, aff=affirmative, np=noun phrase, vb=verbal chunk, vbneg=negated verbal chunk, adjp=adjectival phrase, advp=adverbial phrase, pp=prepositional phrase.

[5] POS abbreviations: PRP=personal pronoun, AUX=auxiliary verb, VB=main verb (non-inflected), PREP=preposition, PRPA=personal pronoun in accusative, RB=adverb.

[6] Frame abbreviations: INAN=inanimate NP, ANIM=animate NP, VBZ=inflected main verb, IS=is, VBG=gerund, PP=prepositional phrase, TO=to (prep.), ON=on (prep.).

classification	occ. (%)	Comment
correct	21 (21.4%)	at least one reasonable mapping is found
non-mappable	23 (23.5%)	clause consists of noises/hesitations only
ungrammatical	28 (28.6%)	e.g., incomplete phrase, no verb
preprocessing	13 (13.3%)	problem is caused by errors in POS tagger/segmenter/parser
mapper	13 (13.3%)	problem due to incompleteness of mapper

Table 3: Summary of classification results for mapper output

```
 [pred] ([vb_fin] ([aux]  (will/AUX)
                   [head] (talk/VB))
         [pp_obj] ([prep]     (to/PREP)
                   [theme/an] (you/PRPA)))
[modif]   (again/RB)
```

Since chunks like `advp` or `conj` are not part of the WordNet frames, we remove these from the parsed chunk sequence, before a mapping attempt is being made.[7]

In our example, WordNet yields 14 frames for 6 senses of the main verb `talk`. The mapper already finds a "perfect match"[8] for the first, i.e., the most frequent sense[9] of the verb (mapping 4 can be estimated to be more accurate than mapping 3 since also the preposition matches to the input string). This will be also the default sense to choose, unless there is a word sense disambiguating module available that strongly favors a less frequent sense.

Since WordNet 1.5 does not provide detailed semantic frame information but only general subcategorization with extensions such as "animate/inanimate", we plan to extend this information by processing machine-readable dictionaries which provide a richer set of semantic role information of verbal heads.[10]

It is interesting to see that even at this early stage of our project the results of this shallow analysis are quite encouraging. If we remove those clauses from the test set which either should not or cannot be mapped in the first place (because they are either not containing any structure ("non-mappable") or are ungrammatical), the remainder of 47 clauses already has a success-rate of 44.7%. Improvements of the system components *before* the mapping stage as well as to the mapper itself will further increase the mapping performance.

[7] These chunks can be easily added to the mapper's output again, as shown in the example.

[8] Partial matches, such as mappings 1 and 2 in this example, are allowed but disfavored to perfect matches.

[9] In WordNet 1.5, the first sense is also supposed to be the most frequent one.

[10] The "agent" and "theme" assignments are currently just defaults for these types of subcat frames.

5 Future Work

It is obvious from our evaluation, that most core components, specifically the mapper need to be improved and refined. As for the mapper, there are issues of constituent coordination, split verbs, infinitival complements, that need to be addressed and properly handled. Also, the "linkage" between main and relative clauses has to be performed such that this information is maintained and not lost due to the segmentation into short clauses.

Experiments with speech recognizer output instead of transcripts will show in how far we still get reasonable frame representations when we are faced with erroneous input in the first place. Specifically, since the mapper relies on the identification of the "head verb", it will be crucial that at least those words are correctly recognized and tagged most of the time.

To further enhance our representation, we could use speech act tags, generated by an automatic speech act classifier (Finke et al., 1998) and attach these to the short clauses.[11]

6 Summary

We have presented a system which is able to build shallow semantic representations for spontaneous speech in unrestricted domains, without the necessity of extensive knowledge engineering.

Initial experiments demonstrate that this approach is feasible in principle. However, more work to improve the major components is needed to reach a more reliable and valid output.

The potentials of this approach for NLP applications that use speech as their input are obvious: semantic representations can enhance almost all tasks that so far have either been restricted to narrow domains or were mainly using word-level representations, such as text summarization, information retrieval, or shallow machine translation.

[11] Sometimes, the speech acts will span more than one short clause but as long as the *turn*-boundaries are fixed for both our system and the speech act classifier, the re-combination of short clauses can be done straightforwardly.

7 Acknowledgements

The author wants to thank Marsal Gavaldà, Mirella Lapata, and the three anonymous reviewers for valuable comments on this paper.

This work was funded in part by grants of the Verbmobil project of the Federal Republic of Germany, ATR – Interpreting Telecommunications Research Laboratories of Japan, and the US Department of Defense.

References

Steven Abney. 1996. Partial parsing via finite-state cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, Prague, Czech Republic*, pages 8–15.

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. In *Proceeedings of AAAI-94*.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ANLP Conference, Washington DC*, pages 24–29.

Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries and Torsten Zeppenfeld. 1997. The Janus-RTk SWITCHBOARD/CALLHOME 1997 Evaluation System. In *Proceedings of LVCSR Hub5-e Workshop, May 13-15, Baltimore, Maryland.*

Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel and Klaus Zechner. 1998. CLARITY: Inferring Discourse Structure from Speech. In *Proceedings of the AAAI 98 Spring Symposium: Applying Machine Learning to Discourse Processing, Stanford, CA*, pages 25-32

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the ICASSP-92*, volume 1, pages 517–520.

Alon Lavie. 1996. *GLR*: A Robust Grammar-Focused Parser for Spontaneously Spoken Language.* Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceeedings of the 31th Annual Meeting of the ACL*, pages 235–242.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Five papers on WordNet. Technical report, Princeton University, CSL, revised version, August.

Piek Vossen, Pedro Diez-Orzas, and Wim Peters. 1997. The Multilingual Design of EuroWordNet. In *Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997*

Alex Waibel, Michael Finke, Donna Gates, Marsal Gavaldà, Thomas Kemp, Alon Lavie, Lori Levin, Martin Maier, Laura Mayfield, Arthur McNair, Ivica Rogina, Kaori Shima, Tilo Sloboda, Monika Woszczyna, Torsten Zeppenfeld, and Puming Zhan. 1996. JANUS-II - advances in speech recognition. In *Proceedings of the ICASSP-96.*

Wayne Ward. 1991. Understanding spontaneous speech: The PHOENIX system. In *Proceedings of ICASSP-91*, pages 365–367.

Klaus Zechner. 1997. Building chunk level representations for spontaneous speech in unrestricted domains: The CHUNKY system and its application to reranking Nbest lists of a speech recognizer. M.S. Project Report, CMU, Department of Philosophy. Available from `http://www.contrib.andrew.cmu.edu/~zechner/publications.html`