GSLT 2001
Speech Technology 1
Term paper

Per-Anders Jande
TMH, KTH
jande@speech.kth.se

Linguistic Knowledge in High Level Automatic Speech Processing

1. Introduction

The term *high level automatic speech processing* is in this paper used as a comprehensive term for processing systems aimed at automatically converting speech to some non-speech higher level (i.e. meaningful, linguistically or otherwise) representation.

The aim of this paper is to discuss how the integrated use of more linguistic knowledge can facilitate efficiency and accuracy of high level automatic speech processing. Some examples of work in this field are briefly presented and discussed. Further, the main concern of this paper is not the practical methods associated with high level speech processing, but rather what types of linguistic and phonetic knowledge can be used in speech processing and for what purposes. In short, one could say that this paper focuses mainly on *what* rather than on *how*.

There are basically two different high level automatic speech processing paradigms, represented by *automatic speech recognition* (ASR) systems and *automatic speech understanding* (ASU) systems, respectively. The two types of speech processing systems can be said to reflect different paradigms in the sense that they have fundamentally different system-internal assumptions about what speech is. This will be discussed more extensively in section 2 of this paper.

There are many ASR systems for commercial use and/or research and development purposes. These systems are used e.g. for automatic dictation, in dialogue systems and for voice control in handicap aids. To the extent ASU is used (e.g. in dialogue systems), it is mostly as a linguistic interpretation back-end extension of an ASR system and not as an integrated perception/interpretation system.

The paper is organised in the following way: first, the two paradigms for high level speech processing (represented by ASR and ASU, respectively) are presented and their theoretical difference discussed. Then, the reminder of the paper is devoted to discussing the integration of high level linguistic and phonetic-prosodic information into high level automatic speech processing systems.

2. Two Paradigms for High Level Automatic Speech Processing

As mentioned, there are two main paradigms in automatic speech processing, differing mainly in the system-internal assumption about what speech is. That is, we have two paradigms for how to define *speech*. The paradigm defining speech in ASR systems is concerned with the linguistic code and views speech simply as a sequence of units. The paradigm defining speech in ASU systems is concerned with the meaning conveyed by the speech signal and views the speech signal as a carrier of a message. Thus, an ASR system and an ASU system have – at least from a linguist's point of view – very different goals. An ASR system is to produce a written text as output from spoken language input and does not take on the task of unravelling the message conveyed by the speech signal and its context. An ASU system, on the other hand, is focused on extracting the user's intentions and finding the user's intended meaning by *interpreting* spoken utterances. In this interpretation process, determining the exact word sequences in utterances is not necessarily important. Although often treated as such in practice, ASU is thus not a mere elaboration of ASR, but should be seen as something inherently different.

2.1 Automatic Speech Recognition

The units to be recognised in ASR are typically words as represented in a lexicon. State-of-the-art ASR systems typically take speaker-independent spontaneous speech as input and are set to recognise a large number of word units (tens of thousands for speaker adaptive systems and up to around thousand for non-adaptive systems) present in the system's lexicon. Automatic speech recognition systems are not concerned with the "meaning" of the spoken utterances it processes and performance measures depend on the share of correctly recognised words, not considering if the words are important for the general semantic meaning of an utterance/a series of utterances.

There are a number of prototypical practical approaches to automatic speech recognition. These are commonly grouped into four types: template-based approaches, knowledge-based approaches, stochastic approaches (mainly using hidden markov models, HMMs, and Viterbi type search algorithms) and connectionist approaches (using artificial neural networks). Stochastic models using HMMs are predominant in state-of-the-art commercial systems (cf. e.g. Blomberg & Elenius, 2000; Waibel & Lee, 1990).

As mentioned, ASR systems typically use very simple language models, in which language is seen merely as a statistically constrained set of strings of words. The probability of a word being situated at a certain position in a string of words is dependent on the *n-1* preceding words in the string. A typical language model is the *trigram* model. In a trigram model, *n=3* (cf. e.g. Jelinek, 1999). Taking more words into consideration generally does not improve recognition accuracy very much, but does increase the processing load significantly. The probabilities used by the trigram models are derived through training on a corpus and adjusted with the use of smoothing techniques to handle words not found in the training data.

Trigram models are statistically rather than linguistically motivated[1], although they of course catch regularities in syntax, semantics and pragmatics – especially for languages with relatively fixed word order (cf. e.g. Jelinek, 1991). The trigram models thus do their work quite well, although more linguistic knowledge is being implemented in ASR systems to improve accuracy and processing time.

2.2 Automatic Speech Understanding

A widely used approach to ASU is to extract the meaning of a spoken utterance in a post-processing stage after an initial ASR component has done its job. However, much information that can be used to decrease errors in pure word determination can come from higher level linguistic and pragmalinguistic knowledge. This suggests that applying word recognition and message interpretation in sequence is not optimal.

Also, even if a front-end ASR component correctly converts the speech signal into text, much information is lost since meaning is largely represented in fundamentally different ways in text and in speech. Text is usually more formal and more "literal" and thus very much dependent on lexical and grammatical means of conveying information. Speech has more dimensions of coding the information, e.g. the opportunity of using prosody (and gestures, in face-to-face communication), deictic expressions etc. A speaker also normally has access to immediate feedback and is thus less dependent on formal aspects such as using specific terms and "correct" and full sentences; if the listener has understood the message, there is no point in elaborating. Anyone who has tried to read transcribed spoken conversations knows that such texts are not always easy to understand. This is due to the different situations in which the two language media are used. Text is typically non-direct and monological, while speech is direct, dialogical and has access to situational context.

The output of an ASU system is some sort of semantic representation of the speaker's spoken message. However, this representation is only a formal semantic representation strongly correlated with linguistic units, i.e. a "literal", language based semantic representation. Representations of multi-purpose utterances, fuzzy-purpose utterances (e.g. utterances from strictly social speech where the content is not as important as the fact that there *is* communication) and utterances where "reading between the lines" is necessary to get the meaning are not possible using such solutions to semantic representation. This although the kind of utterances just described constitute the larger part of everyday speech. On the other hand, applications are usually not constructed for social small-talk (although polite phrases and such will have to be

[1]It can be argued that human language knowledge is based on statistics derived from exposure to language. As an extension of this view, one could argue that the best way to teach a machine to analyse language is to let it learn from data. In principle, I agree with this view. However, it does not mean that linguistic knowledge (or "metalinguistic" knowledge, i.e. the knowledge possessed by the linguist rather than the language user) cannot be useful when accomplishing this task.

recognised as “litter”) and the speech input can be assumed to have a reasonably literal meaning, provided that the speaker uses the system in a serious manner.

The semantic representation output from an ASU system is typically used to execute a query or a command in a dialogue system or as input to a natural language generator in an automatic speech translation system. In a dialogue system, the semantic representation can also be used to facilitate the subsequent interaction with the user.

ASU systems are generally not evaluated in terms of correctly recognised words (although the ASU system may produce a written text representation of the input utterance at some stage of processing). Instead, ASU systems can be evaluated in terms of the share of correctly executed commands or queries or correct translations (for ASU components implemented in specific applications) or in terms of the share of correct semantic representations as judged by humans.

3. Integrating Linguistic and Phonetic Knowledge

To illustrate the importance of higher level linguistic information in the interpretation of a spoken message, let us consider human speech interpretation for a while. It is reasonable to assume that a person – when listening to speech under normal circumstances – has the main goal of determining the meaning of an utterance or a series of utterances and only under very special circumstances to determine the exact word sequences the utterance consists of. It is, further, much easier to determine which sequence of words is uttered if you already have some idea about what the meaning of the utterance “should” be. The situational and linguistic/pragmalinguistic context usually provides a good basis to form hypotheses at general and more specific levels about what the speaker is going to say. Also, it is much easier to determine the phoneme status of a segment of speech if you know the word or word sequence in which the segment occurred (or if you have a *hypothesis* about what the word/sequence is). In fact, a speech segment cannot be said to have a phoneme status outside the context of some meaningful utterance (note that even a one-phoneme utterance is associated with a lot of pragmatic information such as world knowledge and situation context).

In a similar way as humans use high level linguistic knowledge, such knowledge can also facilitate automatic speech processing. Local knowledge that can be derived from the signal (e.g. conveyed by prosody) can be used directly in interpreting the speech signal to reduce the search space and perplexity and thus increase both processing speed and accuracy. Higher level semantic and pragmatic knowledge can be accumulated over a discourse and be used for top-down perplexity reduction in the further processing of speech input.

The introduction of linguistic knowledge in a system is often conceived of as hand-typed rules of different sorts. Here, however, what is meant by linguistic knowledge is rather a specification of *what* to do and not necessarily *how* to do it. That is, linguistic knowledge can be used to specify the goals of a system although it does not determine exactly how to achieve these goals. Phonetic and linguistic knowledge at a low level is

used in all ASR and ASU systems in order to reduce linguistically unimportant variability and to hypothesise word strings in a more efficient manner. Low level linguistic knowledge (in some sense) can be said to be used already at the signal pre-processing level for input signal quantisation. However, there is much more linguistic knowledge than phonetic/ phonotactic detail and syntax. Below, some examples of work on integrating high level linguistic and phonetic-prosodic knowledge into high level automatic speech processing systems are presented and discussed.

3.1 Higher Level Linguistic Information

As mentioned, linguistic knowledge is often used at the back end of a speech recogniser to correct errors. However, there are systems that use high-level linguistic knowledge sources for top-down (hypothesisation) purposes (cf. e.g. Young et al., 1989; Young, 1990).

Young et al. (1989) describe a dialogue system which uses semantics, pragmatics, dialogue information and user domain knowledge to predict what the user is likely to say next. The predictions are used to compile grammars specific for the predicted input, prior to the speech processing stage. This both increases accuracy and shortens processing time when the predictions are correct. The predictions are made at several levels of specificity. The most specific predictions are checked first. If the speech input does not match any utterance allowed by the most specific prediction-based grammar (i.e., matching scores fall beneath a threshold value), the system re-processes the signal using a less specific grammar. Since the more general grammar allows more alternative utterances, the search space and perplexity is larger and processing time increases at the same time as accuracy decreases. However, the system can still handle unexpected input. Eventually – if no good-enough match is found – the system reaches a stage of prediction constraints which is identical to the constraints of the full possible grammar of the system. All dialogue considerations are thus ignored at this stage.

The system's predictions are based on dialogue considerations such as user goals, user plans and dialogue foci. The grammars used are semantic network grammars. These have constraints depending on semantic meaningfulness as well as on syntactic well-formedness when allowing or disallowing a certain utterance hypothesis. The dynamically produced prediction-specific semantic grammars use only a subpart of the system lexicon. The grammars are attached to hierarchical goal trees (reflecting tentative intended goals for – i.e., purposes of – the utterance) with subgoals as nodes and domain concepts as leafs. The domain concepts are the concepts that can be involved in trying to achieve the subgoal of the node. These concepts are the objects that the system is able to talk about and attributes associated with the objects. The domain concepts are not independent objects but represented in a structure that enables inheritance and multiple relations between objects. A user model is used to relate different goal nodes to each other using knowledge about what the user knows and is presently able to infer. Each domain concept leaf has a pre-compiled semantic network grammar attached to it. The

grammar of each domain concept is thus pre-compiled, while the full grammar associated with the goal tree is dynamically composed from the part-grammars. The semantic net grammars have word categories as end nodes. A normal HMM-based speech recogniser with phonemes and words as recognition units – constrained by the grammar given by the current level of specificity at the current dialogue stage – is used to produce scores for the match between the speech input and allowed utterances under the given constraints.

The domain concepts and their relations, the possible dialogue goals and the semantic network grammars used in the system were all compiled by hand. The possibility to compile these components by hand is clearly dependent on a very limited domain. For larger domains, automatic methods for doing this will have to be developed. The task of defining a system of high level knowledge becomes increasingly difficult the less constrained the domain. That is, there is not only an exponential growth in labour as more concepts are added. The complexity of the relations also increases. For the extreme case of an unconstrained domain, it is simply not possible to use a framework like that described in Young et al. (1989), irrespective of the access to automatic methods. However, unconstrained domain speech understanding would require true artificial intelligence, which seems to be a long way into the future and is not within the scope of this paper.

The matching in the system described by Young et al. (1989) is made on the word level with the language model as a front-end connecting meaning representations to word sequences. However, with the help of prosodic cues, the matching between input and hypothesised utterances could be done (at least in part) on a higher level and the higher level linguistic knowledge would thus truly integrated into the speech processing system.

3.2 Prosodic Information

Prosodic information is typically not used in commercial ASR/ASU systems. However, much progress has been made in the last few years in making use of prosodic information in experimental systems. Prosody has been used for different purposes and at different levels of analysis. Waibel (1987), for example, shows that prosodic knowledge sources (syllable duration, the relations between the duration of voiceless and vocalic segments and the syllable duration, word intensity patterns and word stress patterns) conveys much additional information for word hypothesisation (in relation to segmental phonetic knowledge sources).

Further, Lee et al. (2001) have used prosody at a higher level to localise prosodic phrase boundaries prior to word hypothesisation. The detection of prosodic phrase boundaries is useful since many phonological rules – especially for coarticulation – are constrained to work within a prosodic phrase. Correct prosodic boundary detection can thus be used to increase word recognition accuracy. Also, the use of prosodic boundary information can reduce the search space (cf. Lee et al. 2001).

Shriberg & Stolcke (2001) present a list of applications for which they have used macro-level and word-level prosody. These applications include sentence segmentation, disfluency detection, topic segmentation, modelling of turn-taking including different types of overlapping speech and also word recognition (like Waibel, 1987). For deriving prosodic features, they used a framework that learns "raw" prosodic features (durational and pitch-based) from speech data with aligned phonetic transcriptions. The learning procedure does thus not utilise any a priori phonological prosodic categories (such as e.g. dialogue acts and phrase boundaries). Such categories are instead derived from the learned "raw" classes.

As a final example of the use of prosodic information in speech processing systems, prosody is widely used in the Verbmobil speech-to-speech translation system (Wahlser, 2000; Niemann et al., 1997; Hess et al., 1996). The areas in which prosody is used as an information source in Verbmobil are parsing (Kompe et al., 1997), exclusion of erroneous semantic representations, dialogue act segmentation and classification (Warnke et al., 1997) as well as translation, generation and speech synthesis (which are areas outside the scope of this paper).

The type of prosodic information most widely used in Verbmobil is information associated with clause boundaries. Such information is used both for parsing and for dialogue processing (Niemann et al., 1997). In addition to boundary detection on the macrosyntactic level, Hess et al. (1996) also mention the use of prosodic information in disambiguation at many levels (semantic and pragmatic disambiguation, sentence mood determination, compound versus non-compound disambiguation) and search space reduction, e.g. through the use of stress patterns (at the word level).

In the Verbmobil prosodic module, prosodic information is derived from the speech independently from word hypothesisation. Word hypothesis information and prosodic information are then combined before decisions about word sequences are made. Prosody is thus treated as a knowledge source independent from segmental phonetic information. Further, prosody is seen as additional side information (Hess et al., 1996).

Since prosodic phrase boundaries are correlated with syntactic boundaries, macro-level prosodic bottom-up information can be used in combination with higher level top-down linguistic information. This is, however, not done in any systematic way in the systems just described.

4. Conclusions

Two paradigms for high level automatic speech processing, differing in their view of what speech is, have been presented. Automatic speech recognition and automatic speech understanding systems have been compared to illustrate the difference between the paradigms. The benefits of using more linguistic and phonetic knowledge in high level automatic speech processing have been discussed. Some literature showing that such knowledge can indeed increase accuracy and reduce search space in ASR and ASU systems have been presented. It has been argued that it is not optimal to use different

knowledge sources independently and that there is a need for integrating knowledge sources in the automatic interpretation of spoken language.

References

Blomberg, M.; K. Elenius (2000) "Automatisk igenkänning av tal" ('Automatic recognition of speech', in Swedish) *Institutionen för tal, musik och hörsel, KTH*

Hess, W.; A. Batliner; A. Kießling; R. Kompe; E. Nöth; A. Petzold, M. Reyelt; V. Strom (1996) "Prosodic Modules for Speech Recognition and Understanding in VERBMOBIL" in Y. Sagisaka, N. Campbell & N. Higuchi (eds.) *Computing Prosody, Approaches to a computational analysis and modelling of the prosody of spontaneous speech, Springer Verlag*

Jelinek, F (1991) "Up from trigrams! The struggle for improved language models" *Proceedings of Eurospeech 91 3:1037-1040*

Jelinek, F. (1999) "Statistical methods for speech recognition" *MIT Press, Cambridge, Massachusetts*

Kompe, R.; A. Kießling; H. Niemann; E. Nöth; A. Batliner; S. Schachtl; T. Ruland; H.U. Block (1997) "Improving parsing of spontaneous speech with the help of prosodic boundaries" *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), II:811-814*

Lee, S.; K. Hirose; N. Minematsu (2001) "Incorporation of prosodic modules for large vocabulary continuous speech recognition" *Proceedings of the 2001 International Speech Communication Association (ISCA) Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*

Niemann, H; E. Nöth; A. Kießling; R. Kompe; A. Batliner (1997) "Prosodic processing and its use in Verbmobil" *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), I:75-78*

Shriberg, E.; A. Stolcke (2001) "Prosody modelling for automatic speech understanding: An overview of recent research at SRI" *Proceedings of the 2001 International Speech Communication Association (ISCA) Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*

Wahlster, W. (2000) "Mobile speech-to-speech translation of spontaneous dialogs: an overview of the final Verbmobil system" in Wahlster, W. (ed.) *Verbmobil: Foundations of Speech-to-Speech Translation, Springer Verlag*

Warnke, V.; R. Kompe; H. Niemann; E. Nöth (1997) "Integrated dialog act segmentation and classification using prosodic features and language models" *Proceedings of Eurospeech 97 1:207-210*

Waibel, A. (1987) "Prosodic knowledge sources for word hypothesization in a continuous speech recognition system" *Proceedings of the 1987 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2:856-859*

Waibel, A.; Lee, K. (1990) "Readings in speech recognition" *Morgan Kaufmann Publishers, inc. San Mateo, California*

Young, S.R. (1990) “Use of dialogue, pragmatics and semantics to enhance speech recognition” *Speech Communication 9:551-564*

Young, S.R.; A.G. Hauptmann; W.H. Ward; E.T. Smith; P. Werner (1989) “High level knowledge sources in usable speech recognition systems” *Communications of the ACM (Association for Computing Machinery) 32(2)*