Anaphora Resolution in a speech recognition environment

Dimitri Woei-A-Jin, Leon Rothkrantz
Philips Research,PFL-Aachen
Data and Knowledge engineering,
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
L.J.M.Rothkrantz@its.tudelft.nl

Abstract

In this paper the methods applied to resolve anaphora in the speech recognition application scenario are described. In these applications shallow-parsing has been used, which provides no information about sentence structure and only relevant phrases are returned. As syntactic information is an essential input to most resolution algorithms, a new method had to be developed. It is based on the model presented in [Str98]. In addition a set of filters is employed to overcome the lack of syntactic information so that it is possible to determine some of the dependencies between different phrases, which are needed to successfully solve anaphora.

Introduction

In the past few years, the Man-machine Interface group of Philips Research in Aachen has built an Electronic Programming Guide (EPG) with a multi-modal interface for television program recording within the Speech Interface for Consumer Electronics project (SPICE) [Kel00]. The SPICE-EPG allows input from two different modalities. It is possible to do all tasks hands-free with spoken input only, or control the device with touch screen input in addition to speech. The current prototype consists mainly of visual feedback for displaying information (e.g. a list of program items matching the user's selection) and spoken output to guide the user through the dialogue.
The user can select a set of programs from the database by specifying one or more of the following items in one utterance: date, time, genre, channel, title, or description. Currently the SPICE-EPG lacks a method to solve references to previously mentioned topics or objects (anaphora). Since anaphora very often abound in naturally occurring discourse, they are a critical part of spontaneous speech understanding. It is therefore important to be able to solve anaphora in a user friendly speech environment.
At Delft University of Technology there is a project running on multimodal interfaces [Rot00]. Systems have now been developed for several modalities. At this moment a prototype of a system has been developed which integrates speech, lipreading and recognition of facial expressions. For years there is a project running on the development of an information system on timetables of trains based on automatic speech recognition (ALARON). A corpus of 5.000 annotated dialogues has been created. Anaphora occurs frequently in these dialogues.
Even though much research already has been done on anaphora resolution, it is not as simple as picking a suitable anaphora resolution model, implement it and have a working anaphora resolution module. Most anaphora resolution models are very theoretic and many assumptions about available information to solve anaphora are usually not present in a speech understanding application. The SPICE-EPG system uses a shallow parser based on a stochastic context-free grammar to parse the user utterance. This parser provides no syntactic information. No information about sentence structure is generated either. The parser only extracts those words or phrases that are meaningful with respect to the current task. The rules specified in the grammar do not have to cover the complete user utterance, but only the meaningful phrases in the utterance must be represented. These are called concepts. The so-called filler model allows the handling of meaningless phrases, which cannot be assigned to a concept [Aus95] [Sou00].
The main reasons for this model are that most of the input sentences from the user are not grammatically correct according to textual grammar rules, and the speech recognizer introduces additional errors which lead to parsing problems for the used grammars. Most anaphora resolution models depend on the syntactic information which are generated by these grammars, but are not available in the SPICE-EPG application. It is therefore important that a method is found to compensate for this lack of information.

Related work

The most simple anaphora resolution model is based on the recency constraint alone, which states that the antecedent of an anaphor should be the most recently mentioned object that satisfies all the constraints [All95]. This model does not need any syntactic or semantic information to determine which referent is the most preferred. However this model has a very low performance of 47%, tested on texts from several magazines and part of a story [Str98].
A more sophisticated model to determine the preferred reference is the centering model by [Bre87]. This model is based on the intuition that most discourses are organized around an object that the discourse is about, which is called the center. If any object in the local context is referred to by a pronoun in the current sentence, then the center of that sentence must also be pronominalized. The center must be the most preferred discourse entity in the local context that is referred to by a pronoun. Continuing with the same center from one sentence to the next is preferred over changing the center. The preference for a center is determined by the function of the phrases in the text: subject first, direct object next, indirect object, and then the other discourse entities in the sentence. The centering model has good performance of 72,9% and 76,0% for a variant of the centering model, which were tested on texts from several magazines and part of a story [Str98]. However, the centering model requires information like: subject, direct object, and indirect object to determine the preferences for the centers, which are not available with the shallow parser used for the SPICE-EPG.
An alternative to centering is *Never look back*, which is a model proposed by [Str98]. This model is based on the notion that the preference for referents can be determined by the attentional state of the hearer which is determined by how new or old an entity is in the discourse and to the hearer. Discourse entities are grouped into three categories: hearer-old discourse entities (OLD), mediated discourse entities (MED) and hearer-new discourse entities (NEW). Hearer-old discourse entities are entities which are already in the discourse model of the hearer, mediated discourse entities are linked to entities already in the discourse model, and hearer-new discourse entities are not yet in the discourse model. There is a preference relation among these entities to which references will refer, which are stated in table 1, where $x < y$ indicates a preference for x.

Table 1. Precedence of entities.

if $(x \in \text{OLD} \wedge y \in \text{MED}) \vee$ $(x \in \text{OLD} \wedge y \in \text{NEW}) \vee$ $(x \in \text{MED} \wedge y \in \text{NEW})$ **then** $x < y$
if $(x, y \in \text{OLD} \vee x, y \in \text{MED} \vee$ $x, y \in \text{NEW})$ **then** **if** $(\text{utt}_x < \text{utt}_y)$ **then** $y < x$ **if** $(\text{utt}_x = \text{utt}_y \wedge \text{pos}_x < \text{pos}_y)$ **then** $x < y$

This model performs better than the centering model, with a performance of 85,3%, tested on texts from several magazines and part of a story [Str98], and needs less complex information.
Other methods often employed to solve references are heuristics, where a set of experimental rules are determined by trial and error experiments. The performance for different heuristic methods ranges from 69,1% to 92,4% on a variety of corpera [Byr99], [Ken96], [Mar00], [Mcc96], [Mur96]. Often syntactic and semantic information are needed, and rules are often application specific.

Anaphora Resolution Model

The anaphora resolution model implemented during this project is based on 'Never look back', since 'Never look back' has good performance to determine the preferred referent, without the need of complex syntactic information, which is not available with the shallow parser.
There are some small differences with the proposed model, because the applications differ. In [Str98], the entities in the salience list (the list with the discourse entities which are currently in focus ordered by their attentional state) will be removed if they are not used in the utterance. In the implemented model entities are not removed during the system turn, only during the user turn. This is done because the user is not really in a dialogue with the system, since the output is presented on the display most of the time, rather than using speech to inform the user. The user can ignore the system output and refer back to what he said earlier.

However, even though no syntactic information is necessary to determine the preferred referent, still some information about dependencies are needed between several phrases, to determine which referents can and which referents cannot be referred to considering the context of a sentence. Take for example *Please record the one at eight o' clock*. According to the model of Never look back, the attentional state of the hearer may cause a preference among discourse entities in the following order: *CNN, world news*, and *displayed programs*. The word *record* indicates however that *the one* should refer to a program, so *CNN* is not an optional referent. The hearer can infer from the syntax that *record* applies to *the one* and not for example to *eight o' clock*. To determine how different phrases relate to each other, three general solutions for three different cases are formulated for use in an environment where no syntactic information is available to determine dependencies.

In the first case a concept modifies another, but they do not necessarily follow each other, so it is possible that other concepts are placed between them. For example *Please record for me the one at eight o'clock.* To determine whether *record* applies to *the one*, a set of premisses are matched with the properties of *the one*. The premisses which are matched in this case are for example: The target should not be a person, not be abstract, and should not be a different type than a program. The number of premisses may vary each concept, depending on the concepts it will modify, but in general the number is around 3 – 5 premisses. If these premises hold, it can be assumed that *record* applies to *the one* and that *the one* is a program.

In the second case a concept modifies another, and they always follow each other according to a set pattern. For example: *the eight p.m. news or the news at eight p.m.* In the case the concepts are treated as a single concept by the grammar. A filter will find this construction and creates a concept for *eight p.m.* and assigns it as a subconcept of *the news*. Later during reference handling, the reference resolution modules knows that *eight p.m.* modifies *the news* and can assign the constraint *start time = eight p.m.*

The third case is a more difficult form of reference, where a concept refers to a property of a concept, which is a reference itself. For example *the second program of the previous list* or *from the last list, the second program from below*. Usually these concepts occur together, so a similar approach as in the second case can be used. The two concepts are treated as a single concept by the grammar. A filter will find this construction and creates a concept for the last list and assigns it as a superconcept of the second program from below. Later during reference handling, the reference resolution module solves the superconcept first, and looks in this superconcept to find the referent of the concept.

Evaluation

This reference resolution module was tested offline and online on a limited number of examples. Unfortunately no suitable corpus was available for testing in the SPICE-EPG application. Corpera previously used to test the SPICE-EPG lacked any form of references, and no suitable corpus was found. It has not been tested on corpora used by others, because the reference solution module is not a stand alone program, but is designed to be an integrated part of a system used to create speech understanding consumer applications. For this a corpus is needed which suits such an application. Because no such corpus was available a set of examples of references which the system should ideally be able to handle given by co-workers were used to test the system offline on a set of tasks which were used to the usability of the SPICE-EPG.

The offline test consisted of 35 references, of which 33 were solved correctly. None of these errors could be attributed to the lack of information about dependencies between the different concepts, but rather to the lack of information present in the system. In *the serial I saw last night, when will it be continued*, both *the serial I saw last night* and *it* were correctly resolved. This was because no information is present about which serial was seen last night. Therefor no referent was found for the definite description. Because of this, the system had trouble placing it in the salience list, so that it could not be correctly resolved to *the serial I saw last night.*

The online test can be considered as an evaluation of the whole system with the reference module. The test consisted of nine tasks where queries with references were used to achieve certain end results, like searching for a specific program and adding it to the record list. In the online tests reference resolution is much more difficult, because recognition errors introduce concepts which are not I the utterance. Also less complex sentences are used, because this increases the chance that the system does understand what is said. The speech recognizer has trouble recognizing certain words, of which some are important for reference resolution.

Words like *he, she*, and the definite article the are often not recognized. Another problem is that certain concepts which can be referred to are badly recognized. This is the case with name of actors and other persons. For instance whether a movie with Robert de Niro would be running this night, was not recognized in the twenty attempts made, and other actors fared little better. In the few cases were the actor was indeed recognized, the reference *he* or *she* was not or misrecognized, causing the actor to leave the focus of

attention of the reference resolution module. It is not possible to solve this problem in the reference resolution module directly, because the module has no way to determine whether a shift in focus was due to misrecognition of the user or because the user shifted his focus. Therefore pronominal anaphora resolution to persons is not tested in the online evaluation. However, considering the fact that these kind of references were successfully resolved on the offline test, it is expected to work correctly should the referent and reference be recognized.
Sometimes during online testing when a misrecognition occurs, a reference is wrongfully introduced (usually the pronouns *it* and its or the demonstratives *that, this* and *them*). When this happens three things may occur: the reference is resolved to anything, the reference is resolved to a concept currently in the focus of the user, or a concept out of the focus of the user.
It was found however that recognition errors do not really create strange shifts in the focus of attention of the system, which would cause correctly recognized references to be resolved wrongly. During the tests, some misrecognitions contained references which were resolved to the concept in focus, so no shifts in focus were occurred. Also when the system would move away from the desired task, for example display a totally different topic, the user will try to move back to the task at hand, rather than just relentlessly trying to have the system recognize the utterance.
Finally we tested the system offline on the corpus of 5.000 human-human dialogues. The performance was very good. More then 95% of the reference handling was successful. But we to realize that the different types of anaphora are simple cases.

Conclusion

In this paper it is described what methods were employed to solve anaphora in a speech understanding environment for applications where no syntactic information is present. To determine the preference for possible referents, the model proposed by [STR98} is implemented. This model is based on the assumption that the hearer's attentional state is based on how new the information is in relation to the hearer and the discourse. This is a simple model which does not rely on syntactic information about the concepts in the sentence. To compensate for the lack of syntactic information, three general solutions are proposed to determine the dependencies between two concepts:
The first one is to look at the properties of the target concept, and match them with a set of premises stated by the source concept. If these premises hold, it is assumed that the source concept modifies the target concept, and additional constraints can be added. This method is best used when two concepts do not necessarily follow each other directly, and it is possible to have other 'non-related' concepts between them.
The second method is to have the grammar treat the concepts as a single concept. A filter later extracts the two concepts, and assigns the concept which modifies the other as a subconcept of the other one. This is especially useful when the two concepts always follow each other directly. It is easier to determine the relation this way, and misassignments are less likely to occur.
The third method is used in a special case, where a so called compound reference occurs. A reference refers to a property of another concept, which is a reference itself. Usually these concepts occur directly after each other, so a similar approach can be used as the second one. The concepts are grouped together by the grammar and a filter extracts the different concepts, and assigns the concept which contains the property which the other concept refers to as the superconcept of the other. The superconcept is resolved first, and is used to resolve the other concept.
During evaluation of the system, no resolution errors could be traced to errors in determining the dependencies between the concepts. In the online test, many errors are generated by misrecognition of the user by the system.
Future work could address the tuning of the entire system to be more robust for recognition errors and the resolution of one anaphora, which is also dependent on syntactic information.

Acknowledgements

The authors are grateful to Petra Philips, Thomas Portele and Andras Kellner for their helpand advice. The first part of the work was done at Philips Research

References

[All95] J.F. Allen. "Natural Language Understanding – 2nd Edition," The Benjamin/Cummings Publishing Company, 1995.
[Aus95] H. Aust, M. Oerder, F. Seide, and V. Steinbiss. The Philips automatic train timetable information system. Speech Communication, 17(3-4):249-262, Nov. 1995.
[Bre87] S.E. Brennan *et al* ."A centering approach to pronouns," *Proceedings of the Association for Computational Linguistics*, pp. 155-162. July 1987.

[Byr99] D.K. Byron. Resolving Pronominal Reference to Abstract Entities. Ph.D. thesis, University of Rochester Department of Computer Science. June 1999.

[Kel00] A. Kellner *et al.* “SPICE - a first research prototype. Conversational User Interfaces,” *PFL-Aachen Report 1463/00*, July 2000.

[Ken96] C. Kennedy, B. Boguraev - Anaphora for everyone: pronominal anaphora resolution without a parser. Proceedings of the 16th International Conference on Computational Linguistics COLING'96, Copenhagen, Denmark, 5-9 August 1996

[Mar00] P. Martínez-Barco, & M. Palomar. Dialogue structure influence over anaphora resolution. In *MICAI 2000: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence, vol.1793*. O. Cairo, L.E. Sucar and F.J. Cantú Eds. Acapulco (Mexico). Springer-Verlag. pp. 515-525. ISBN: 3-540-67354-7, April 2000.

[Mcc96] J. McCarthy. A Trainable Approach to Coreference Resolution for Information Extraction. University of Massachusetts Amherst Department of Computer Science. September 2000.

[Mur96] M. Murata. Anaphora Resolution in Japanese Sentences Using Surface Expressions and Examples. Ph.D. thesis, Kyoto University. December 1996.

[Rot00] L.J.M.Rothkrantz *et al.*"Dialogue control in the Alparaon system"*'Proc. Of the third Workshop on Text, Speech and Dialogue*, Brno, Czech Republic. Springer Verlag, 2000, pp 333-338

[Sou00] B. Souvignier, A. Kellner, B. Reuber, H. Schramm & F. Seide. The thoughtful elephant: Strategies for spoken dialogue systems. IEEE Transactions on Speech and Audio Processing, 8(1):51-62, January 2000.

[Str98] M. Strube. “Never Look Back: An Alternative to Centering”, *Proceedings of ACL-98*, pages 1251-1257, 1998.