Co-reference in Japanese Task-oriented Dialogues: A Contribution to the Development of Language-specific and Language-general Annotation Schemes and Resources

Felix Sasaki, Andreas Witt

Universität Bielefeld
Fakultät für Linguistik und Literaturwissenschaft
– Computerlinguistik und Texttechnologie –
{felix.sasaki, andreas.witt}@uni-bielefeld.de

Abstract

This paper describes a corpus of Japanese task-oriented dialogues, i.e. its data, annotations, analysis methodology and preliminary results for the modeling of co-referential phenomena. Current corpus-based approaches to co-reference concentrate on textual data from English or other European languages. Hence, the emerging language-general models of co-reference miss input from dialogue data of non-European languages. We aim to fill this gap and contribute to a model of co-reference on various language-specific and language-general levels.

1. Introduction

The use of standardized markup languages and markup vocabularies offers a lot of advantages for the annotation, archiving and transfer of language data. But markup languages and vocabularies do, in addition, support the process of information modeling. The problem is that markup vocabularies are specific to certain theories and to a certain language or at least a certain language family. This leads to problems in the standardization of linguistic annotations. Ide and Romary, 2003 present the distinction of a general annotation format, a so-called *Virtual Annotation Markup Language* (VAML) and a (theory-, language-, domain-) specific annotation format, a *Concrete Annotation Markup Language* (CAML) as a solution to this problem. In this paper we show, how this division of concrete and virtual markup languages improves the process of choosing a markup vocabulary for annotating a phenomenon. The phenomenon 'co-reference', which is expressed in different languages, i.e. English and Japanese, with different linguistic means, serves as our use case.

2. Language data and data analysis

2.1. A corpus of multiple annotations for various existing annotation schemes

We use the Japanese part of the *tinkertoy corpus* which is described in Sasaki et al., 2002. The corpus is based upon a task-oriented dialogue scenario with two participants: The instructor has pictures of objects, which have to be explained to the constructor, who has the respective building parts. The corpus consists of six dialogues with 2160 utterances. The dialogues are tagged on the morpho-syntactic level at first automatically (Matsumoto et al., 2000) with later manual correction. To test the methodology, one dialogue has been annotated on 25 annotation levels with about 9.000 annotation units, making use of various existing annotation schemes for Japanese morpho-syntactic and semantic categories. These schemes encompass for example the EDR-scheme (EDR, 2002) for the semantic subcategorization of lexical units, the morphosyntactic tagset developed within the Verbmobil project (Kawata and Bartels, 2000) and the co-reference annotation scheme developed by (Kawahara et al., 2002). This large number of concrete, existing annotation schemes is used to evaluate their usability as an input for general, virtual markup languages. All annotation schemes have been used as the basis for existing corpora, which can be used for the evaluation as well. The general idea is not to develop yet another CAML, but to integrate the existing CAMLs into the emerging VAMLs.

2.2. Analysis methodology

Ide and Romary, 2003, claim that "the annotator, must choose a *data architecture* for the primary text and its annotations, which dictates whether annotations are interspersed throughout the document containing the primary text or stored in one or more additional documents linked to the primary text." In our project[1] we use a third *data architecture*: we annotate the same textual resource several times. This annotation technique results in a set of annotated XML-instances differing only in the markup, i.e. the elements, attributes and attribute-values. Because the textual content of all layers is identical, the text can serve as a link between these layers, cf. Witt, 2002. In a document annotated with this multilayer approach several relations between elements can be found, e.g. the relation *identity* which holds if two different elements contain the same range of text. Figure 1 depicts the possible relations (see also Bayerl et al., 2003).

Using this technique, our corpus is used for analyzing the relations between the CAMLs described above on an empirical basis. For the analysis of the configurations between the annotations on different layers, a special-purpose tool has been developed. This inference tool is implemented in Prolog. It allows for the inference of relations for every type of annotation unit, compared to other CAMLs or proposed VAMLs.

[1] Project A2/Sekimo, DFG-Forschergruppe 437/Texttechnologische Informationsmodellierung

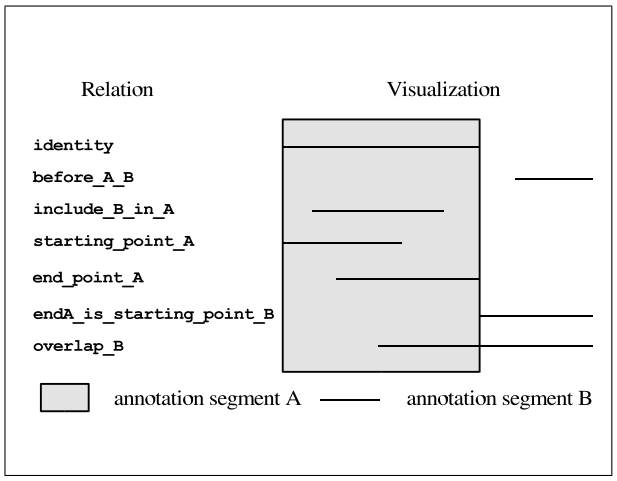

Figure 1: Possible relations between ranges of elements within two documents containing the same data

3. CAMLs and VAMLs for a model of co-reference in Japanese task-oriented dialogues

In this section, various VAMLs and CAMLs for co-referential phenomena in Japanese are introduced. To be able to compare the VAMLs to similiar work described within the project COMOn-REFs (see Vieira et al., 2003), we concentrate on a category which plays an important role for the respective algorithm, namely *definiteness*. Afterwards, we discuss interactional properties of co-referential phenomena which are specific to Japanese dialogues. The creation of all VAMLs depends on the description of morpho-syntactic patterns, i.e. *means*, which have a certain semantic or pragmatic *function*. A formal basis for the description of such patterns can be found in Frajzyngier and Mycielski, 1998. Various VAMLs for Japanese and language-general VAMLs for co-reference are developed. In a conclusion, the relation of existing CAMLs for Japanese to these VAMLs is discussed.

3.1. Corpus-based modeling of co-reference: Current approaches

To be able to compare algorithms for anaphora resolution and to reuse corpus data for training, corpus-based models of co-reference and related phenomena gain more and more importance. Current approaches concentrate mainly on English. Due to this fact, recent projects try to adapt the respective algorithms to other languages, for example the project COMMOn-REFs mentioned above. This project focuses upon Portuguese and French, but also aims at a language-general, corpus-based model for co-reference. Nevertheless, most approaches towards corpus-based models of co-reference concentrate on textual data and European languages. This leads to a gap in the area of spoken dialogues and non-European languages. There has been corpus-based research using English dialogue data (Strube and Müller, 2003). Approaches concerning Japanese (Kawahara et al., 2002) concentrate on textual data. This paper aims at filling this gap with the corpus of Japanese task-oriented dialogues, annotated on several language- and domain specific CAMLs, relevant for co-reference.

3.2. VAMLs for Definiteness

Definiteness is an important category for the developement of algorithms for anaphor resolution. For example, in Vieira and Poesio, 2000, an algorithm which makes use of so-called *Definite Descriptions* (DD, noun-phrases starting with the definite article), is introduced. The algorithm has been developed for English and tested within English corpus data. To be able to reuse the algorithm for other languages, its application to French and Portugese is currently being tested in the project COMMOn-REFs (see Vieira et al., 2002). It seems to be possible to adapt the algorithm to these two languages with an affordable amount of work.

Nevertheless the algorithm depends on the existence of definiteness as a morphosyntactic category, which is expressed e.g. via the definite article. In Japanese, definiteness does not exist as a morphosyntactic category. Also, there is no other morpho-syntactic means which can be used to create a suitable CAML for Japanese, i.e. which can be interpreted directly as a marker for definiteness. To be able to test the useability of algorithms for DD, *intermediate VAMLs* have to be created. They encompass not single morpho-syntactic categories, but morpho-syntactic patterns which have a function similiar to definiteness. To exemplify our methodology, here we focus upon a category called *numeral classifier* (NC). NCs are nominal suffixes which are used in conjunction with numerals. They express semantic characteristics of discourse entities, for example spatial properties. An example of NCs taken from our corpus is given below.

(1) ***kiiroi*** ***bou*** *ga* *ari* *masu* *[...]* ***sore***
yellow stick - exist - [...] it
ADJ N SUBJ V V-AUX [...] PRON
wo *kadokko* *ni* *sashite* *kudasai*
- corner - put please
OBJ N LOC V V-IMP
ex-6-utt3-5, Speaker 1 (Instructor): “There are **yellow sticks** [...] Put **them** into the corner.”

(2) *[...]* ***yon*** *pon* *tomo* *desu* *ka*
[...] four - - - -
[...] four NC PART V-AUX QP
ex-6-utt14, Speaker 2 (Constructor): “All **the four of them** (the sticks)?”

In example 1, a referent ‘kiiroi bou’ “yellow stick” is introduced. A co-referential relation to this referent is expressed via the pronoun ‘sore’ “it” in a subsequent utterance by the same speaker. In example 2, the other speaker refers to the same referent, through the use of a NC: The NC ‘hon’ or its allomorph ‘pon’ expresses the property of being round and long, which is shared by the previously mentioned discourse entity ‘stick’.

The general role of NCs in a model of co-reference has been discussed previously (Sasaki et al., 2002). For the

creation of VAMLs, a more detailed model is necessary. It relies on patterns of morpho-syntactic linguistic means, which are used to express a specific type of (co-)reference. The pattern of NCs in example 2 is the most common in our data: The NC is used in a fragment, without a syntactically related noun. This pattern is used to express discourse-old information, i.e. co-reference to an entity introduced in the discourse before. Other morpho-syntactic patterns are described by Downing, 1996, p. 225: pre-nominal, appositive, summative appositive and Q-float. The pre-nominal pattern is very common in Japanese texts. It is used to express discourse new referents, or to express a co-referential relation to a referent introduced before. In the latter case, the co-referential relation can be classifed as a subtype of *bridging*: A subset relation holds between the first mention and the second mention. We assume that the pre-nominal pattern has the same role in dialogue data. This allows for the creation of several VAMLs which are visualized below.

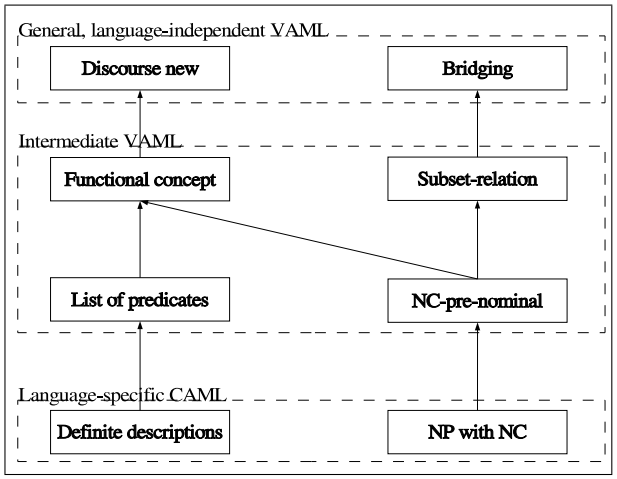

Figure 2: VAMLs and CAMLs for the distinction between discourse new and discourse old entities via numeral classifiers or definite descriptions

In the lower part of Figure 2, language specific, morpho-syntactic categories are introduced as part of language-specific CAMLs, i.e. definite descriptions for English, French and Portugese, and NPs with NCs for Japanese. In the middle, intermediate VAMLs are visualized. To interrelate definite descriptions with NCs, an abstract category *functional concept* is introduced. It has been developed by Löbner, 1985 and is also used within the algorithm for DD to detect new referents within discourse. For English, a list of predicates like "the best" within the NP is applied to trigger the interpretation as a functional concept. For Japanese, the pattern pre-nominal is used as a similar trigger. Since this pattern can also be used to create a bridging relation, it has to be tested whether the same NC has been used in the discourse before. In this case, a subset relation holds between the previous mention and the pre-nominal pattern. The test consists of the application of the predicate `before_B_A`. If before annotation `A` (annotation of the NP with NC) there is the annotation `B` (annotation of the same NC), the result of the test is true. Our data architecture allows for such tests in general, i.e. without knowing the - hierarchical - structure of the corpus in detail.

Although this example of relations between VAMLs and CAMLs seems to be rather procedural, it is not. In our approach, tests as described above and all relations between VAML and CAML are described in a *declarative format*. This format allows for the validation of the relations, i.e. the execution of the tests within annotated data. It makes use of a specific query approach which is described by Sasaki et al., 2004.

3.3. VAMLs for interactional properties of Japanese dialogues

A property of Japanese dialogues is the high frequency of back-channel signals. Not only, but especially due to these signals the definition of the communicative unit *utterance* in Japanese is a difficult task. Example 1 contains two utterances by the same speaker, which are only interrupted shortly by the other speaker. To differentiate various types of co-reference specific to dialogue, the interruption signals have to be interpreted as signals for back-channel or as a meaningful contribution by the other speaker. We make use of patterns within the description of VAMLs for this purpose. The patterns encompass typical lexical units and their morpho-syntactic patterns which can be interpreted as back-channel signal. Although back-channel is a phenomenon which also relies on prosodic features, we do not take such features into account. The reason is that the syntactic features convey enough information to disambiguate back-channel phenomena, cf. Koiso et al., 1998. With this analysis, we are able to create language-specific and language-general VAMLs for the interactional properties of co-reference, see Figure 3 below.

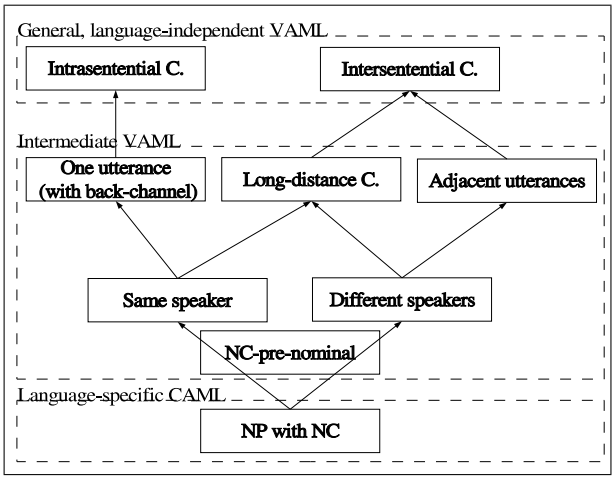

Figure 3: Description of the interactional properties of co-reference on a language-independent and language-specific scale

The verification of the VAMLs and CAMLs in Figure 3 dependents on annotations on several layers, encompassing information on the speaker of the utterance, the distance between referent and antecedent and the role of back-channel signals. On the level of CAMLs, again we focus upon NPs with NCs. Such NPs can be uttered by the same speaker that made the utterance containing the antecedent,

or by the other speaker. The co-referential units might be in different utterances (*intersentential co-reference*) or in the same utterance. In the latter case the role of back-channel becomes important. Morpho-syntactic patterns for back-channel phenomena in the respective VAMLs have to be used to ensure that intrasentential co-reference is not misinterpreted as intersentential co-reference.

An important aspect in Figure 3 is that the VAMLs allow for the integration of the VAMLs for definiteness. For this purpose, the VAMLs for interactional properties might not be based upon the CAML `NP with NC`, but on the VAML `NC-pre-nominal`. Again, whether such an integration is suitable or not depends on the domain in question.

3.4. Applicability of existing CAMLs for the integration into VAMLs

Since our studies on VAMLs for co-reference concentrated on a singular category, namely numeral classifier, nearly all of the existing CAMLs mentioned before can be used to create the VAMLs. In other words, all of the existing annotations which rely on these CAMLs can be integrated into our framework, presupposing that they are converted into the data architecture described above.

Still, for some tasks certain CAMLs are more useful than others. For example the EDR-scheme supplies a detailed, semantic sub-categorization schema of nominal units; this could be used to automatically infer the NCs which might be used with the noun in question. On the other hand, the co-reference scheme developed by (Kawahara et al., 2002) allows for a sub-classification of co-referential phenomena which is language-specific to Japanese, but it encompasses no detailed sub-categorization of lexical units. Such differences stress the need not to rely on a singular CAML, but to combine them, in order to relate them to the VAMLs envisaged. Again, for this purpose it is useful to rely on the data architecture discribed above which allows for such a combination.

4. Summary

This paper described modeling of co-referential phenomena in Japanese task-oriented dialogues. To be able to separate language-general and language- or domain-specific models of co-reference, we rely on a certain data architecture and a query concept which are described elsewhere. Because we concentrate upon a construction dialogue scenario, NCs are very common in the corpus data. Nevertheless for other domains the intermediate VAMLs concentrating on NCs might not be useful. The same holds for the VAMLs which describe interactional properties of Japanese. As stated before, intermediate VAMLs in general depend on the domain in question. So it is necessary to analyze the feasability of the VAMLs empirically. For this purpose, the data architecture and the query approach mentioned above (see (Sasaki et al., 2004)), is used.

5. References

Bayerl, P. S., H. Lüngen, D. Goecke, A. Witt, and D. Naber, 2003. Methods for the Semantic Analysis of Document Markup. In *Proceedings of the 2003 ACM Symposium on Document Engineering*. ACM Press.

Downing, P., 1996. *Numeral Classifier Systems : The Case of Japanese*. Amsterdam: Benjamins.

EDR, 2002. Electronic Dictionary Version 2.0. Technical report, Communications Research Laboratory, Tokyo.

Frajzyngier, Z. and J. Mycielski, 1998. On Some Fundamental Concepts of Mathematical Linguistics: Means and Functional Domains. In C. Martin-Vide (ed.), *Mathematical and Computational Analysis of Natural Language*. Amsterdam: Benjamins, pages 295–310.

Ide, N. and R. Romary, 2003. Encoding Syntactic Annotation. In A. Abeillé (ed.), *Building and Using Parsed Corpora*. pages 281–96.

Kawahara, D., S. Kurohashi, and K. Hashida, 2002. Construction of a Japanese Relevance-tagged Corpus. In *Proceedings of LREC 2002*. Las Palmas.

Kawata, Y. and J. Bartels, 2000. Stylebook for the Japanese Treebank in VERBMOBIL. Technical report, Verbmobil.

Koiso, H., A. Shimojima, and Y. Katagiri, 1998. Collaborative Signaling of Informational Structures by Dynamic Speech Rate. *Language and Speech*, (3-4):323–350.

Löbner, S., 1985. Definites. *Journal of Semantics*, 4:279–326.

Matsumoto, Y., A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara, 2000. *Morphological Analysis System ChaSen version 2.2.1 Manual*. Http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1.pdf.

Sasaki, F., C. Wegener, A. Witt, D. Metzing, and J. Pönninghaus, 2002. Co-reference Annotation and Resources: A Multilingual Corpus of Typologically Diverse Languages. In *Proceedings of LREC 2002*. Las Palmas, Spain.

Sasaki, F., A. Witt, D. Gibbon, and T. Trippel, 2004. Concept-based queries: Combining and Reusing Linguistic Corpus Formats and Query Languages. In *Proceedings of LREC 2004*. Lisabon.

Strube, M. and C. Müller, 2003. A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue. In *Proceedings of ACL-2003*. Sapporo, Japan.

Vieira, R., C. Gasperin, R. Goulart, and S. Salmon-Alt, 2003. From Concrete to Virtual Markup-Language: the Case of COMMOn-REFs. In *Proceedings of the ACL-2003 Workshop on Linguistic Annotation: Getting the Model Right*. Sapporo, Japan.

Vieira, R. and M. Poesio, 2000. An Empirically-based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4):525–579.

Vieira, R., S. Salmon-Alt, and E. Schang, 2002. Multilingual Corpora Annotation for Processing Definite Descriptions. In *Portugal for Natural Language Processing - PorTAL 2002*. Universidade do Algarve, Faro, Portugal.

Witt, A., 2002. Meaning and interpretation of concurrent markup. In *Proceedings of ALLC / ACH 2002*. Tübingen, Germany.