Enhancing Anaphora Resolution for Czech

Vašek Němčík

NLP Laboratory,
Faculty of Informatics, Masaryk University,
Brno, Czech Republic
xnemcik@fi.muni.cz

Abstract. Resolution of anaphoric reference is one of the most important challenges in natural language processing (NLP). Functionality of most NLP systems crucially relies on an accurate mechanism for determining which expressions in the input refer to the same entity in the real world. The immense complexity of this issue has led the research community to adopt predominantly knowledge-poor methods, despite the fact that these are known to be incapable of solving this task reliably. This paper suggests several ways of extending such methods by further resources and mechanisms in order to arrive at a more adequate anaphora resolution procedure.

1 Introduction

Anaphora has been one of the most intensively studied issues in the linguistic research over the past decades. It has been studied from many different perspectives – from the point of view of syntax, semantics, pragmatics, psycholinguistics, computational linguistics, rhetoric, logic, philosophy, etc. Nevertheless, we still seem to be left in the dark about many important aspects of anaphora.

This situation is well apparent from the recent implementations of anaphora resolution (henceforth AR) systems. Since the mid-1990s, most of the implementations have been based on knowledge-poor and machine learning (ML) approaches, relying solely on low-level features such as morphological tags and shallow syntactic labels.[1] This trend is motivated by practical reasons. The individual low-level features can be computed automatically, efficiently and with sufficient accuracy. In contrast, higher-level information, such as a full syntactic parse or underlying semantics, unfortunately can't be obtained reliably enough, and the consequent errors undermine the AR performance considerably.

Knowledge-poor systems have proven themselves as a sensible trade-off between accuracy and computational feasibility. On the other hand, higher-level information is known to play an important role in anaphoric relations and thus can't be ignored altogether. This can be illustrated by the following

[1] The most influential knowledge-poor systems are systems presented by Lappin and Leass ([7]), Kennedy and Boguraev ([5]), Baldwin ([1]), and Mitkov et al. ([12]). ML-based systems will be discussed in section 3.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007, pp. 57–62, 2007. © Masaryk University, Brno 2007

examples that demonstrate the necessity of semantic information and world knowledge for proper treatment of anaphora:

(1) a. After the bartender$_i$ served the patron$_j$, he$_i$ got a big tip.
b. After the bartender$_i$ served the patron$_j$, he$_j$ left a big tip.

(2) If the baby$_i$ does not thrive on raw milk$_j$, boil it$_{*i,j}$.

Although it is obvious that obtaining and combining all types of information relevant to AR is well beyond the scope of today's science, it is worthwhile to use at least certain types of higher-level information. This paper proposes how this can be done for Czech.

Next section suggests how to take advantage of certain more sophisticated linguistic resources to improve the performance of AR. Further, section 3 suggests several ways of adapting AR methods based on machine learning so that they grasp the properties of anaphoric relations in a more plausible way.

2 Exploiting Linguistic Resources for AR

This section gives a number of hints how to extend the common knowledge-poor systems by considering various kinds of higher-level information. Of course, the possibilities depend on what resources are available for the language in question. As this article concerns anaphora resolution with regard to Czech, it reflects particular resources available for Czech. Nonetheless, the following ideas can be straightforwardly applied to any language for which similar resources exist.

To my knowledge, at the moment, there are three AR systems for Czech. The first one is the modular system proposed by Němčík ([13]), encompassing selected salience-based algorithms. The other two systems were presented by Linh ([9]) – one of them is rule-based and the other is based on machine learning. All of these systems take advantage of solely knowledge-poor features and can be straightforwardly extended to use further resources.

The desired extension would ideally help to rule out semantically implausible antecedent candidates that would get otherwise incorrectly chosen by the original system.[2] Not necessarily all such antecedents need to be ruled out, on the other hand, it is important that the enhancing mechanism in question be sound, i.e. it shouldn't rule out correct antecedents.

The first potentially useful resource available for Czech is the Czech Wordnet, described by Pala and Smrž ([15]).[3] Its English version is often used to determine semantic plausibility when dealing with coreference resolution. However, on its own, it is rather useless for resolving pronominal anaphora.

[2] This idea has already been mentioned by Hobbs ([4]).

[3] Strube and Ponzetto ([17]) argue that for practical purposes Wikipedia is a more useful resource, because it doesn't suffer from problems of hand-crafted taxonomies and contains information not only about classes but also idividual real-world instances. Moreover, it is larger and grows faster than Wordnet.

Another type of resource that could be used within the AR process are valency lexicons. For Czech, two of them are available, Vallex and Verbalex.[4] In my opinion, especially Verbalex is very helpful in this cause because its valency slots are annotated with semantic constraints. These are marked using Wordnet synsets, meaning that each slot can be filled only by an concept that is a hyponym of the synset indicated. This can be straightforwadly used in combination with Wordnet as a semantic plausibility check for AR illustrated by the following schema:

(3) a. $Verb_1\ \alpha_1\ \dots\ \alpha_{i-1}\ Y\ \alpha_{i+1}\ \dots$
...
b. $Verb_2\ \beta_1\ \dots\ \beta_{j-1}\ X\ \beta_{j+1}\ \dots$

Let us assume that X is an anaphor and Y an antecedent candidate preceding it.[5] Should Y be a plausible antecedent for X, it should meet the restrictions posed on the valency slot of X. In particular, it should be a hyponym of the synset associated with this valency slot. This mechanism can contribute to the correct resolution of anaphors in the following examples:[6]

(4) a. ***Obsluhující robot***$_i$ odnesl prázdnou misku od ***ovoce***$_j$
Robot (MASC.SG.) took the empty bowl of fruits (NEUT.SG.)
"The robot took away the empty fruit bowl"

b. a Alvar si teprve díky tomu uvědomil,
and Alvar only then thanks to this realized,
"and only after noticing this Alvar realized"

c. že ***ho***$_{*i,j}$ vůbec snědl.
that him/it (MASC./NEUT.SG.) actually ate.
"that he actually ate it."

(5) a. Dolehl k ***němu***$_i$ ***zvuk***$_j$
echoed to him (MASC.SG.) sound (MASC.SG.)
melodického smíchu
of melodic laughter
"A sound of melodic laughter echoed to him"

b. a $\emptyset_{i,*j,*k}$ na okamžik si myslel,
and [he (MASC.SG.)] for a moment thought,
"and for a moment he thought"

c. že je to Mary.
that is it Mary.
"it was Mary."

[4] For more information about Vallex and Verbalex please refer to Lopatková et al. ([10]), and Hlaváčková and Horák ([3]), respectively.

[5] It may well be that $Verb_1 = Verb_2$, that is, that the anaphor and the antecedent are in the same clause.

[6] The examples are for the sake of brevity slightly abridged sentences taken from The Czech National Corpus ([18]).

Obviously, this mechanism is not applicable to all anaphor–antecedent candidate pairs of this kind. The potential hindrances are many – it is not possible to reliably assign a unique valency frame to every sentence, to disambiguate every relevant word and match it with the correct Wordnet synset, and most importantly, neither Wordnet nor Verbalex can cover all words. However, to obtain a similar effect with higher recall, we can engage methods for determining semantic relatedness.

Recently, many interesting corpus-based methods have been proposed that make it possible to measure semantic similarity between words. For instance, Lin ([8]) has formulated a similarity measure based on mutual information between words.[7] A similar measure is adopted in the Sketch Engine tool (Kilgarriff et al., [6]) and can be utilized to approximate suitability of verb–argument combinations. This allows making a more sophisticated choice among top antecedent candidates. As a result, many resolution errors can be avoided, especially in cases when there is only a small difference in salience among top antecedent candidates.

The above-mentioned mechanisms seem to be a very promising first step in integrating semantics into AR systems. Investigation of their potential in practice is subject of my future work.

3 Anaphora Resolution and Machine Learning

This section suggests how AR approaches based on ML can be altered to more closely reflect the properties of anaphoric reference.

Presently, methods based on ML form an integral part of the mainstream AR research. Nevertheless, ML methods are not directly applicable to the AR task, because its structure is unsuitable and it needs to be transformed first to fit the ML concept. To my knowledge, two notable re-formulations of AR as a classification task have been proposed. They are in turn sketched by the following schemata:[8]

(6) Antecedent$_1$ Antecedent$_2$ Anaphor **1/2**

(7) Antecedent Anaphor **Y/N**

Connolly et al. ([2]) suggested instances consisting of an anaphor and two antecedent candidates, the target information left to be learnt being which of these two candidates is "better" for the anaphor in question. This information could be then utilized by a step-by-step elimination of the less plausible candidates to determine the correct antecedent.

The other formulation of the task has been proposed by McCarthy and Lehnert ([11]) and has been used by most of the state-of-the-art systems as the standard one. It postulates instances formed by an anaphor–antecedent candidate pair together with the information whether the candidate is a

[7] The mutual information scores have been computed based on dependency triples extracted from a large parsed corpus.

[8] A word represents a set of features (of the entity hinted by its meaning), symbols in bold represent possible values of the target feature.

valid antecedent of the anaphor or not. This attribute determines whether the instance is understood as positive or negative.

Most ML-based AR systems use knowledge-poor features to describe the individual instances. Unsurprisingly, this poses problems similar to the ones described in the previous section. In my opinion, an important additional problem is that the features are considered out of context. The individual instances provide a very detailed description of the relationship between the anaphor and the antecedent, which is very advantageous for nominal coreference resolution, where the relation between the referred entities plays a more important role than context. However, this view of the task is very unsuitable for grasping pronominal anaphora, where different types of information, such as salience or interplay with other antecedent candidates, play an important role. Moreover, this seems to be yet a bigger issue for Czech, where, compared to English, information structure is not as tightly connected with the syntactic structure of the sentence.

One solution to this problem is introducing new features reflecting salience. In this respect, Ng and Cardie ([14]) have used the result of a syntactic search AR algorithm as a binary feature, and Preiss ([16]) has engaged the salience factors proposed in the rule-based system of Kennedy and Boguraev ([5]). The latter is a very plausible approach, for Czech with a big potential of benefiting from the rich interaction between syntax and information structure. Moreover, I would suggest re-computing the salience model iteratively during the classification phase to account for the information in already resolved links.

Another solution to this problem can be possibly obtained by a different formulation of AR as a classification task. It can be argued that the AR task has inherently the following structure:

(8) Antecedent$_n$ . . . Antecedent$_1$ Anaphor **1/.../n**

Nevertheless, this concept is not very suitable for ML in this form. The main problems lie in data sparseness and the correct linearization of the antecedent candidates – these can be arbitrarily embedded into each other. On the other hand, this formulation of the AR task contains more information about the relevant context, and the information corresponding to the target feature is actually the piece of information we aim to learn – which antecedent to choose for a given anaphor from a list of candidates. The potential of this AR task reformulation needs to be investigated empirically.

4 Conclusion

In this paper, I have discussed the most notable limitation of most state-of-the-art AR systems – the fact that they disregard higher-level cues, even though these are known to play an important role. I have proposed possible ways of taking advantage of higher-level information available in the AR process, namely considering verbal valency constraints and predicate-arguments statistics. I have also suggested several ways of adapting the ML-based AR methods in order to account for the structure of the AR task more closely.

Acknowledgments

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536.

References

1. Baldwin, B.: Cogniac: High precision coreference with limited knowledge and linguistic resources. In: Proceedings of the ACL '97/EACL '97 workshop on Operational factors in practical, robust anaphora resolution. (1997).
2. Connolly, D., Burger, J.D., Day, D.S.: A machine learning approach to anaphoric reference. In: Proceedings of the International Conference on New Methods in Language Processing (NeMLaP), ACL (1994).
3. Hlaváčková, D., Horák, A.: Verbalex – new comprehensive lexicon of verb valencies for Czech. In: Computer Treatment of Slavic and East European Languages, Bratislava, Slovakia, Slovenský národný korpus (2006) 107–115.
4. Hobbs, J.R.: Resolving pronoun references. In Grosz, B.J., Spärck-Jones, K., Webber, B.L., eds.: Readings in Natural Language Processing. Morgan Kaufmann Publishers, Los Altos (1978) 339–352.
5. Kennedy, C., Boguraev, B.: Anaphora for everyone: pronominal anaphora resolution without a parser. In: Proceedings of the 16th conference on Computational linguistics, Morristown, NJ, USA, ACL (1996) 113–118.
6. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The sketch engine. In: Proceedings of the Eleventh EURALEX International Congress. (2004) 105–116.
7. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computatinal Linguistics **20**(4) (1994) 535–561.
8. Lin, D.: Automatic retrieval and clustering of similar words. In: COLING-ACL. (1998) 768–774.
9. Linh, N.G.: Návrh souboru pravidel pro analýzu anafor v českém jazyce. Master's thesis, Charles University, Faculty of Mathematics and Physics, Prague (2006).
10. Lopatková, M., Žabokrtský, Z., Benešová, V.: Valency lexicon of czech verbs VALLEX 2.0. Technical Report 34, UFAL MFF UK (2006).
11. McCarthy, J.F., Lehnert, W.G.: Using decision trees for coreference resolution. In: Proceedings of the 14th International Conference on Artificial Intelligence IJCAI-95, Montreal, Canada (1995) 1050–1055.
12. Mitkov, R., Evans, R., Orăsan, C.: A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In: Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Mexico City, Mexico (February 17–23, 2002).
13. Němčík, V.: Anaphora resolution. Master's thesis, Masaryk University, Faculty of Informatics, Brno (2006).
14. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th Annual Meeting of the ACL. (2002) 104–111.
15. Pala, K., Smrž, P.: Building Czech WordNet. **2004**(7) (2004) 79–88.
16. Preiss, J.: Machine learning for anaphora resolution. Technical Report CS-01-10, University of Sheffield, Sheffield, England (Aug 2001).
17. Strube, M., Ponzetto, S.: WikiRelate! Computing semantic relatedness using Wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), Boston, Mass. (July 2006) 1419–1424.
18. The Czech National Corpus (2006) `http://ucnk.ff.cuni.cz/english/`.