Designing a Language Game for Collecting Coreference Annotation

Barbora Hladká and **Jiří Mírovský** and **Pavel Schlesinger**
Charles University in Prague
Institute of Formal and Applied Linguistics
e-mail: {hladka, mirovsky, schlesinger@ufal.mff.cuni.cz}

Abstract

PlayCoref is a concept of an on-line language game designed to acquire a substantial amount of text data with the coreference annotation. We describe in detail various aspects of the game design and discuss features that affect the quality of the annotation.

1 Introduction

Creating a collection of high quality data is resource-demanding regardless of the area of research and type of the data. This fact has encouraged a formulation of an alternative way of data collection, "Games With a Purpose" methodology (GWAP), (van Ahn and Dabbish, 2008). The GWAP methodology exploits the capacity of Internet users who like to play on-line games. The games are designed to generate data for applications that either have not been implemented yet, or have already been implemented with a performance lower than human. The players work simply by playing the game - the data are generated as a by-product of the game. The more enjoyable the game is, the more users play it and the more data is acquired.

The GWAP methodology was first used for on-line games with images (van Ahn and Dabbish, 2004) and later with tunes (Law et al., 2007),[1] in which the players try to agree on a caption of the image/tune. The popularity of these games is enormous and generates a huge amount of data. *Onto games* (Siorpaes and Hepp, 2008) brought another type of input data to GWAP – video and text.[2]

The situation with text is slightly different. One has to read a text in order to identify its topics. Reading texts takes more time than observing images and the longer text, the worse. Since the game must be of a dynamic character, it is unimaginable that the players would spend minutes reading an input text. Therefore, it must be opened to the players 'part' by 'part'.

So far, besides the Onto games, two more games with texts have appeared: *What did Shannon say?*[3], the goal of which is to help the speech recognizer with difficult-to-recognize words, and *Phrase Detectives*[4] (Kruschwitz, Chamberlain, Poesio, 2009), the goal of which is to identify relationships between words and phrases in a text. No information about their popularity has been published yet.

Motivated by the GWAP portal, the LGame portal[5] dedicated to language games has been established. The LGame portal has been opened with the *Shannon game*, a game of intentionally hidden words in the sentence, where players guess them, and the *Place the Space* game, a game of word segmentation.

2 Coreference

Coreference occurs when several referring expressions in a text refer to the same entity (e.g. person, thing, fact). A *coreferential pair* is marked between subsequent pairs of the referring expressions. A sequence of coreferential pairs referring to the same entity in a text forms a *coreference chain*. The coreferential pairs and the coreference chains cover only the identity relation.

Many projects for various languages on the coreference annotation by linguists are running. The annotated data serve as a basis for further linguistic study of coreference, and most importantly also to train and test procedures for automatic coreference resolution, which is a task that

[1] www.gwap.org

[2] www.ontogame.org

[3] lingo.clsp.jhu.edushannongame.html

[4] www.phrasedetectives.org

[5] www.lgame.cz

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 52–55,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

many other applications can benefit from, e.g. text summarization, question answering, and information retrieval.

Manual annotation is costly and time consuming. We propose a design of the PlayCoref game – to appear at the LGame portal – as an alternative way of the coreference annotation collection, and most importantly, of a substantially larger volume than any expert annotation can ever achieve.

3 The PlayCoref Game

3.1 Game Design

We prepare the game for Czech and English first. However, PlayCoref can be played in any language.

The game is designed for two players. The game starts with several first sentences of the document displayed in the players' sentence window. According to the restrictions put on the members of the coreferential pairs, parts of the text are unlocked (i.e. they are active) while the other parts are locked (i.e. they are inactive); both of them are graphically distinguished. In our case, only nouns and selected pronouns are unlocked. The players mark coreferential pairs between the individual unlocked words in the text (no phrases are allowed). They mark the coreferential pairs as undirected links.

During the session, the number of words the opponent has linked into the coreferential pairs is displayed to the player. The number of sentences with at least one coreferential pair marked by the opponent is displayed to the player as well. Revealing more information about the opponent's actions would affect the independency of the players' decisions.

If the player finishes pairing all the related words in the visible part of the document (visible to him), he asks for the next sentence of the document. It appears at the bottom of the player's sentence window. The player can remove pairs created before at any time and can make new pairs in the sentences read so far. The session goes on this way until the end of the session time. More than one document can be present in the session.

After the session, the players' scores are calculated and displayed.

Instructions for the Players Instructions for the players must be as comprehensible and concise as possible. To mark a coreferential pair, no linguistic knowledge is required, thus no extensive annotation guidelines need to be formulated. It is all about the text comprehension ability.

3.2 Game Data

Any textual data can be used in the game, but the following pre-processing steps are necessary.

Tagging Most importantly, the morphological tagging (usually preceded by tokenization) is required to recognize part-of-speech categories (and sub-part-of-speech categories), in order to lock/unlock individual words for the game. For most languages, tagging is a well solved problem (e.g. for Czech: the MORČE tagger[6], for English: TnT tagger[7]).

Text Parts Locking In the game, we work with coreferential links between the individual words only. The coreferential pairs that link larger text parts consisting of clauses or even several sentences are disregarded. Their marking requires linguistic knowledge and extensive training.

Our research shows that pronouns that are usually members of such "undesirable" links can be detected automatically in advance (at least in Czech). They will get locked, so the players will not consider them at all during the sessions.

Automatic Coreference Resolution According to the way we calculate the players scores (see below), an automatic procedure for coreference resolution is required. If this procedure works on a different layer than the surface layer, further automatic processing of the data may be needed.

4 Data Quality

4.1 Players' Score

We want to obtain a large volume of data so we must first attract the players and motivate them to play the game more and more. As a reward for their effort we present scoring. We hope that the players' appetite to win, to confront with their opponents and to place well in the long-term top scores tables correlates with our research aims and objectives.

Our goal is to ensure the highest quality of the annotation. The scoring function should reflect the game data quality and thus motivate the players to produce the right data. An agreement with

[6] ufal.mff.cuni.cz/morce

[7] www.coli.uni-saarland.de/˜thorsten/tnt/

the manual expert annotation would be a perfect scoring function. But the manual annotation is not available for all languages and above all, it is not our goal to annotate already annotated data.

An automatic coreference resolution procedure serves as a first approximation for the scoring function. Since the procedure does not work for "100%", we need to add another component. We suppose that most of the players will mark the coreferential pairs reliably. Then an agreement between the players' pairs indicates correctness, even if the pair differs from the output of automatic coreference resolution procedure. Therefore, the inter-player agreement will become the second component of the scoring function. To motivate the players to ask for more parts of the text (and not only "tune" links in the initially displayed sentences), the third component of the scoring function will award number of created coreferential links.

The players get points for their coreferential pairs according to the equation $pts_A = w_1 * ICA(A, acr) + w_2 * ICA(A, B) + w_3 * N(A)$ where A and B are the players, acr is an automatic coreference resolution procedure, ICA stands for the inter-coder agreement that we can simultaneously express either by the F-measure or Krippendorff's α (Krippendorf, 2004), N is a contribution of the number of created links, and weights $0 \leq w_1, w_2 \leq 1$, $w_1, w_2, w_3 \in R$ (summing to 1) are set empirically.

The score is calculated at the end of the session and no running score is being presented during the session. From the scientific point of view, the scores serve for the long term quality control of the players' annotation.

4.2 Interactivity Issues

The degree of a player-to-player interactivity contributes to the attractiveness of the game. From the player's point of view, the more interactivity, the better. For example, knowing both his and the opponent's running score would be very stimulating for the mutual competitiveness. From the linguistics' point of view, once any kind of interaction is allowed, statistically pure independency between the players' decisions is lost. A reasonable tradeoff between the interactivity and the independency must be achieved. Interactivity that would lead to cheating and decreasing the quality of the game data must be avoided.

Allowing the players to see their own running score would lead to cheating. The players might adjust their decisions according to the changes in the score. Another possible extension of interactivity that would lead to cheating is highlighting words that the opponent used in the coreferential pairs. The players might then wait for the opponent's choice and again, adjust their decisions accordingly. Such game data would be strongly biased. However, we still believe that a slight idea of what the opponent is doing can boost inter-coder agreement and yet avoid cheating. Revealing the information about the opponent's number of pairs and number of sentences with at least one pair offers not zero but low interactivity, yet it will not harm the quality of the data.

4.3 Post-Processing

The players mark the coreferential links undirected. This strategy differs from the general conception of coreference being understood as either the anaphoric or cataphoric relation depending on the "direction" of the link in the text. We believe that the players will benefit from this simplification and so will the data quality. After the session, the coreference chains are automatically reconstructed from the coreferential pairs.

4.4 Evaluation

Data with manually annotated coreference will be used to measure the game data quality. We will also study how much the scoring function suffers from the difference between the output of the automatic coreference resolution procedure and the manual annotation (gold standard). For Czech, we will use the data from PDT 2.0, for English from MUC-6.

PDT 2.0 [8] contains the annotation of grammatical and pronominal textual coreference. Nominal textual coreference is being annotated in PDT 2.0 in an ongoing project (Nedoluzhko, 2007). Since the PDT 2.0 coreference annotation operates on the so-called tectogrammatical layer (layer of meaning) and PlayCoref plays on the surface layer, the coreferential pairs must be projected to the surface first. The process consists of several steps and only a part of the coreferential pairs is actually projectable to the surface (links between nodes that have no surface counterpart get lost).

[8] ufal.mff.cuni.cz/pdt2.0

MUC-6 [9] operates on the surface layer. This data can be used in a much more straightforward way. The coreferential pairs are marked between nouns, noun phrases, and pronouns and no projection is needed. The links with noun phrases are disregarded.

Evaluation Methods For the game data evaluation, well established methods for calculating an inter-annotator agreement in the coreference annotation will be employed. These methods consider a coreference chain to be a set of words and they measure the agreement on the membership of the individual words in the sets (Passonneau, 2004). Weighted agreement coefficients such as Krippendorf's α (Krippendorf, 2004) need to be used - sets of words can differ only partially, which does not mean a total disagreement.

5 Further Work

Acquisition Evaluation Process The quality of the game annotation undergoes standard evaluation. Apart from collecting, assuming the game reaches sufficient popularity, long-term monitoring of the players' outputs can bring into question new issues concerning the game data quality: How much can we benefit from presenting a document into more sessions? Should we prefer the output of more reliable and experienced players during the evaluation? Should we omit the output of 'not-so-reliable' players?

Named Entity Recognition The step of the named entity recognition will be applied in the subsequent stages of the project. Multi-word expressions that form a named entity (e.g. "Czech National Bank") will be presented to the players as a single unit of annotation. We also plan to implement a GWAP for named entity recognition.

6 Conclusion

We have presented the concept of the PlayCoref game, a proposed language game that brings a novel approach to collecting coreference annotation of texts using the enormous potential of Internet users. We have described the design of the game and discussed the issues of interactivity of the players and measuring the player score – issues that are crucial both for the attractiveness of the game and for the quality of the game data. The game can be applied on any textual data in any language, providing certain basic tools also discussed in the paper exist. The GWAPs are open-ended stories so until the game is released, it is hard to say if the players will find it attractive enough. If so, we hope to collect a large volume of data with coreference annotation at extremely low costs.

Acknowledgments

We gratefully acknowledge the support of the Czech Ministry of Education (grants MSM-0021620838 and LC536), the Czech Grant Agency (grant 405/09/0729), and the Grant Agency of Charles University in Prague (project GAUK 138309).

References

Klaus Krippendorf. 2004. Content Analysis: An Introduction to Its Methodology, second edition, chapter 11, Sage, Thousand Oaks, CA.

Udo Kruschwitz, Jon Chamberlain, Massimo Poesio. 2009. (Linguistic) Science Through Web Collaboration in the ANAWIKI project. In *Proceedings of the WebSci'09: Society On-Line*, Athens, Greece, in press.

Lucie Kučová, Eva Hajičová. 2005. Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of the 5th International Conference on Discourse Anaphora and Anaphor Resolution*, San Miguel, Azores, pp. 97–102.

Edith. L. M. Law et al. 2007. Tagatune: A game for music and sound annotation. In *Proceedings of the Music Information Retrieval Conference*, Austrian Computer Soc., pp. 361–364.

Anna Nedoluzhko. 2007. Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závoslostním korpusu (Annotating extended coreference and bridging relations in PDT). Technical Report, UFAL, MFF UK, Prague, Czech Republic.

Rebecca J. Passonneau. 2004. Computing Reliability for Coreference. *Proceedings of LREC*, vol. 4, pp. 1503–1506, Lisbon.

Katharina Siorpaes and Martin Hepp. 2008. Games with a purpose for the Semantic Web. *IEEE Intelligent Systems Vol. 23, number 3*, pp. 50–60.

Luis van Ahn and Laura Dabbish. 2004. Labelling images with a computer game. In *Proceedings of the SIGHI Conference on Human Factors in Computing Systems, ACM Press*, New York, pp. 319–326.

Luis van Ahn and Laura Dabbish. 2008. Designing Games with a Purpose. *Communications of the ACM*, vol. 51, No. 8, pp. 58–67.

Marc Vilain et al. 1995. A Model-Theoretic Coreference Scoring Scheme. *Proceedings of the Sixth Message Understanding Conference*, pp. 45–52, Columbia, MD.

[9] `cs.nyu.edu/faculty/grishman/muc6.html`