Coreference Annotation of an Arabic Corpus using a Virtual World Game

Wateen Aliady [1,2], Abdulrahman Aloraini[1,3], Christopher Madge[1], Juntao Yu[4], Richard Bartle[4], and Massimo Poesio[1]

[1]Queen Mary University of London, United Kingdom
[2]Imam Mohammad Ibn Saud Islamic University, Saudi Arabia
[3]Qassim University, Saudi Arabia
[4]University of Essex, United Kingdom
{w.a.a.aliady, a.aloraini, c.j.madge, m.poesio}@qmul.ac.uk
{j.yu, rabartle}@essex.ac.uk

Abstract

Coreference resolution is a key aspect of text comprehension, but the size of the available coreference corpora for Arabic is limited in comparison to the size of the corpora for other languages. In this paper we present a Game-With-A-Purpose called *Stroll with a Scroll* created to collect from players coreference annotations for Arabic. The key contribution of this work is the embedding of the annotation task in a virtual world setting, as opposed to the puzzle-type games used in previously proposed Games-With-A-Purpose for coreference.

1 Introduction

Coreference resolution is the task of clustering the mentions in a text that refer to the same real world entity. In the following example of coreference resolution, bold phrases are said to corefer as they point to the same discourse entity, a person named Ibn Sina.

ابن سينا عالم وطبيب اشتهر بالطب والفلسفة.

Ibn Sina is a scientist and doctor who was known for philosophy and medicine.

من أهم أعمال **العلامة** كتاب القانون في الطب.

One of the most famous writings of **the scientist** is The Canon of Medicine.

Coreference resolution is a key element of text comprehension (Poesio et al., 2016; Wu et al., 2021). Identifying references to entities in the context is essential for meaning interpretation. In addition, anaphoric references are an important aspect of textual cohesion, as they connect different parts of the text to ensure its unity. Resolving anaphoric references is essential for most Natural Language Processing (NLP) applications, including automatic translation, information extraction and topic detection (Bouzid and Zribi, 2020).

Collecting coreference annotations from experts can be expensive, so crowdsourcing is often employed (Snow et al., 2008). This can be done using a crowdsourcing platform (Poesio et al., 2008) or by embedding the annotation task in a game in a seamless manner. Such games are referred to as Games-With-A-Purpose (GWAP) (Von Ahn and Dabbish, 2005; Von Ahn, 2006; Von Ahn et al., 2006a,b; Chamberlain et al., 2008; Poesio et al., 2013a; Lafourcade et al., 2015). GWAPs are games designed to collect judgments from players using their gaming skills and language competence; the main reward for players is entertainment. GWAPs have been used e.g., for biological data collection (Kleffner et al., 2017), and, in AI, for image processing (Von Ahn and Dabbish, 2005) and natural language processing (Chamberlain et al., 2008; Krause et al., 2010; Venhuizen et al., 2013; Fort et al., 2014; Dziedzic, 2016; Kicikoglu et al., 2019; Madge et al., 2019b,a; Bonetti and Tonelli, 2020). Using GWAPs for manual annotation is particularly well-suited when the aim is to collect large corpora, that would be too expensive to create using other forms of crowdsourcing (Poesio et al., 2013b).

The objective of this research is to create a GWAP called *Stroll with a Scroll* for Arabic coreference annotation. The motivations for our work are:

- The fact that the available Arabic coreference corpora are limited in size. In the CoNLL-2012 shared task the Arabic portion is about 1/3 of the Chinese and English subsets, comprising about 300k tokens. This is considered a barrier to improving the coreference resolution models accuracy (Pradhan et al., 2012).
- More in general, there is limited work on GWAPs for Arabic language annotation in comparison with English. To our knowledge

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 388 - 393

there is no game with the purpose of collecting Arabic coreference annotation.

Our main contributions are:

- To start a path towards using gamification to attract public engagement to contribute to the creation of larger Arabic coreference corpora, and more in general Arabic NLP corpora;
- The adoption of a virtual world setting, which we expect would increase the chances of attracting players but whose use is still limited in GWAPs for corpus annotation;

2 Related Work

2.1 Games with a Purpose for NLP

The first examples of Games-With-A-Purpose in AI are the well-known *ESP game* for image labelling and other games from Luis von Ahn's lab (Von Ahn and Dabbish, 2005; Von Ahn et al., 2006a,b; Seemakurty et al., 2010). Among the first GWAPs for NLP are *Jeux-de-Mots* for French lexical acquisition (Lafourcade, 2007) and, for English coreference, *Phrase Detectives* (Chamberlain et al., 2008). Other examples are *OnTo-Galaxy* to populate an ontology in English and collect synonyms for German verbs (Krause et al., 2010), *Wordrobe* for English word sense labelling (Venhuizen et al., 2013), *Zombilingo* for French dependency syntax annotation (Fort et al., 2014), *RoboCorp* for Polish named entities annotation (Dziedzic, 2016), *WordClicker* for English part of speech annotation (Madge et al., 2019b), *High School SuperHero* for Italian abusive language annotation (Bonetti and Tonelli, 2020) and *NameThatLanguage* for language recognition (Cieri et al., 2021).

There are some GWAPs for Arabic NLP, including *tashkeelWAP* for digitizing Arabic diacritics (Kassem et al., 2016), *3arosty* for Arabic named entities annotation (Sabty et al., 2016), *3ammeya* to build a corpus for Arabic dialects (Osman et al., 2015) and a GWAP to map Modern Standard Arabic to Arabic regional dialects (Nasser et al., 2013). However, to the best of our knowledge, this is the first GWAP for Arabic coreference and the first GWAP to embed the task of collecting Arabic annotations in a 3D virtual game.

Many of the early NLP GWAPs were essentially gamified versions of annotation tools. Attempts to produce more engaging games include *Puzzle Racer* and *Ka-boom!* for word sense disambiguation (Jurgens and Navigli, 2014). More recently, an engaging game design was developed for *WordClicker*, a part of speech tagging game where players take the role of a baker who fills jars with the part of speech it represents (Madge et al., 2019b).

2.2 GWAPs for Coreference

Phrase Detectives is an online interactive active game to collect English coreference annotation released in 2008. The game has two modes to participate in annotation. The first mode to select a markable that corefers to another highlighted markable and the second mode to validate other players' submitted answers. By 2019, the game had collected over 5 million annotations from more than 50,000 players; the 2nd release of the corpus was the largest crowdsourced corpus for coreference and one of the largest crowdsourced corpora for NLP (Poesio et al., 2019).

Wormingo is an online game to collect English coreference annotation. It creates a novel technique called motivation-annotation paradigm. That highlights the importance of text comprehension in producing accurately coreferenced corpora and making the annotation task easier. Text comprehension is essential in the motivational part of the game that is demonstrated by linguistic puzzles. The annotation part comes after the motivational part and follows the design of *Phrase Detectives* (Kicikoglu et al., 2019).

2.3 GWAPs Embedded in Virtual Worlds

One approach to making games more engaging is to embed them in the virtual world scenarios familiar from most video games. One example of GWAPs adopting this approach is *High School Superhero* (Bonetti and Tonelli, 2020, 2021), a 3D role-playing game is created for abusive language annotation in a sentence level.

Other example is *LingoTowns* (Althani et al., 2022), an isometric world consisting of towns. It hosts three mini-games: *PhraseFarm*, *CafeClicker* and *Lingotoruim*.

The more recent *Borderlands Science* (Waldispühl et al., 2020) is an integration of citizen science game named *Phylo* (Kawrykow et al., 2012) into a massively multiplayer online game called *Borderlands*. In three months, they have collected 50 million puzzle solutions.

3 Stroll with a Scroll

In this paper we introduce *Stroll with a Scroll*, a GWAP for (Arabic) coreference annotation in which the player is an agent embedded in a 3D world.

3.1 Game Design

In *Stroll with a Scroll*, players find themselves in an ancient middle eastern fictional town located in the desert. They roam around this town being represented by an avatar that is dressed in an ancient middle eastern garment as shown in Figure 1.

The game has a treasure hunt theme, with puzzles hidden in the text. To motivate players, we follow the motivation-annotation paradigm introduced by *Wormingo* that uses puzzles and gamification techniques to motivate the players (Kicikoglu et al., 2019). The inclusion of linguistic puzzles increases players' comprehension of text thus, understanding is required to perform coreference annotation.

There are plenty of chests scattered around the town which the player has to find. Only one chest is presented at a given moment; the player is guided to chest location through a navigation system that is displayed on the top right corner presented in Figure 1. The navigation system has three colours: red, yellow, and green to show how far is the avatar from the chest.

The player starts by opening a chest that has a scroll within it. The scroll has textual content with missing parts of information as these pieces were torn because these scrolls are old. The player must guess the lost parts by solving puzzles.

If the player guesses the right word, 10 puzzle points will be added. If a player fails to guess the word, no points will be added.

3.2 Coreference Annotation

The annotation task is presented as questions following the approach used in *Phrase Detectives* (Chamberlain et al., 2008) and *Wormingo* (Kicikoglu et al., 2019). Two types of questions are presented to the player: annotation questions, and validation questions. In the annotation, the player is asked to decide if a mention, colored by red, is discourse new, discourse old (If the player decides that a mention is discourse old, they must select the nearest antecedent from suggested antecedents highlighted in blue) or skip answering as in Figure 2. On the other hand, in the validation, the player is asked to validate the model (Aloraini et al., 2020) or other players.

3.3 Preprocessing

In earlier games such as *Phrase Detectives* , the preprocessing of documents to annotate only involved mention identification (Poesio et al., 2017). However, if a coreference resolver is available, carrying out a preliminary coreference annotation increases

Figure 1: A screenshot of the game.

the potential of a game to collect larger number of annotated documents, as annotation by human players can be driven by uncertainty about the annotation, as in an active learning setting (Li et al., 2020).

The input to *Stroll with a Scroll* is pre-annotated to extract mentions and coreference links using the first neural coreference resolver for Arabic (Aloraini et al., 2020) that achieved higher results than than the existing state-of-the-art system (Björkelund and Kuhn, 2014) on Arabic coreference resolution.

3.4 Aggregation

Stroll with a Scroll follows *Phrase Detectives* (Chamberlain et al., 2008) by using Mention Pair Annotation (Paun et al., 2018) to aggregate user annotations.

4 Discussion

We introduced *Stroll with a Scroll* , a new GWAP for annotating coreference in Arabic. Our GWAP is based on the motivation-annotation paradigm from *Wormingo* in having two disjoint parts: the puzzles part and the annotation part. This division ensures that the orthogonal game design mechanics e.g., aiming, driving and dropping that are the main contributors to most of the popular video games are separated from the annotation task. Video games are separated from the annotation task, so as not to negatively impact the annotation accuracy (Tuite, 2014; Madge et al., 2019a). However, the motivation and the annotation are both embedded in a virtual world scenario: document search involves finding chests in an old town, and filling gaps in the document is naturally presented as reconstructing the scroll. We expect that this novel setting will make the game more attractive to certain types of players who are more interested in 3D games than in puzzles.

5 Limitations

The future based rewards of the annotation part might discourage the players to continue. Furthermore, in the current development stage, the player does not have the option to select antecedents outside of the suggested ones.

6 Conclusion

Games-With-A-Purpose for collecting text annotations are an increasingly popular alternative to crowdsourcing platforms. Even so, to our knowledge there is no GWAP of collecting Arabic coreference annotation. We present a 3D virtual world GWAP of collecting coreference annotations for Arabic corpus. We expect the adoption of a virtual world setting would increase the chances of attracting players.

Figure 2: The annotation task embedded in a 3D game.

References

Abdulrahman Aloraini, Juntao Yu, and Massimo Poesio. 2020. Neural coreference resolution for Arabic. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*.

Fatima Althani, Chris Madge, and Massimo Poesio. 2022. Less text, more visuals: Evaluating the onboarding phase in a GWAP for NLP. In *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*, pages 17–27, Marseille, France. European Language Resources Association.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore, Maryland. Association for Computational Linguistics.

Federico Bonetti and Sara Tonelli. 2020. A 3D role-playing game for abusive language annotation. In *Workshop on Games and Natural Language Processing*, pages 39–43.

Federico Bonetti and Sara Tonelli. 2021. Challenges in designing games with a purpose for abusive language annotation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 60–65.

Saoussen Mathlouthi Bouzid and Chiraz Ben Othmane Zribi. 2020. A generic approach for pronominal anaphora and zero anaphora resolution in arabic language. *Procedia Computer Science*, 176:642–652.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. In *in Proceedings of the International Conference on Semantic Systems (I-Semantics 08)*, pages 42–49.

Christopher Cieri, James Fiumara, and Jonathan Wright. 2021. Using games to augment corpora for language recognition and confusability. In *Proc. of Interspeech: 22nd Annual Conference of the International Speech Communication*.

Dagmara Dziedzic. 2016. Use of the free to play model in games with a purpose: the robocorp game case study. *Bio-Algorithms and Med-Systems*, 12(4):187–197.

Karën Fort, Bruno Guillaume, and Hadrien Chastant. 2014. Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 2–6.

David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464.

Lin Kassem, Caroline Sabty, Nada Sharaf, Menna Bakry, and Slim Abdennadher. 2016. tashkeelwap: A game with a purpose for digitizing arabic diacritics.

Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Luis Sarmenta, Mathieu Blanchette, and Jérôme Waldispühl. 2012. Phylo: A citizen science approach for improving multiple sequence alignment. *PloS one*, 7:e31362.

Doruk Kicikoglu, Richard Bartle, Jon Chamberlain, , and Massimo Poesio. 2019. Wormingo: a'true gamification'approach to anaphoric annotation. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–7.

Robert Kleffner, Jeff Flatten, Andrew Leaver-Fay, David Baker, Justin B Siegel, Firas Khatib, and Seth Cooper. 2017. Foldit standalone: a video game-derived protein structure manipulation interface using rosetta. *Bioinformatics*, 33(17):2765–2767.

Markus Krause, Aneta Takhtamysheva, Marion Wittstock, and Rainer Malaka. 2010. Frontiers of a paradigm: exploring human computation with digital games. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 22–25.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition with the jeuxdemots prototype. In *SNLP'07: 7th international symposium on natural language processing*, page 7.

Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. 2015. *Games with a Purpose (GWAPs)*. Wiley.

Belinda Z. Li, Gabriel Stanovsky, and Luke Zettlemoyer. 2020. Active learning for coreference resolution using discrete annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8320–8331, Online. Association for Computational Linguistics.

Chris Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2019a. The design of a clicker game for text labelling. In *2019 IEEE Conference on Games (CoG)*, pages 1–4. IEEE.

Chris Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2019b. Incremental game mechanics applied to text annotation. In *in Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 545–558.

Sara Nasser, Nada Sharaf, Mohamed Khamis, Slim Abdennadher, and Caroline Sabty. 2013. Collecting arabic dialect variations using games with a purpose: A case study targeting the egyptian dialect. In *Proceedings of the 2nd Workshop on Games and Natural Language Processing (GAMNLP 2013)*.

Maya Osman, Caroline Sabty, Nada Sharaf, and Slim Abdennadher. 2015. Building a corpus for arabic dialects using games with a purpose. In *in 2015 First International Conference on Arabic Computational Linguistics (ACLing), IEEE*, pages 21–25.

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio. 2018. A probabilistic annotation model for crowdsourcing coreference. *http://aclweb. org/anthology/D18-1000*, pages 1926–1937.

Massimo Poesio, Jon Chamberlain, and Udo Kruschwitz. 2017. *Phrase Detectives*, pages 1149–1176.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013a. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1):1–44.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013b. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1):1–44.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, volume 1, pages 1778–1789. Association for Computational Linguistics.

Massimo Poesio, Udo Kruschwitz, and Jon Chamberlain. 2008. ANAWIKI: Creating anaphorically annotated resources through web cooperation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016. *Anaphora resolution*. Springer.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. pages 1–40.

Caroline Sabty, Mirna Yacout, Mohamed Sameh, and Slim Abdennadher. 2016. Gamified collection of arabic named entity recognition data.

Nitin Seemakurty, Jonathan Chu, Luis von Ahn, and Anthony Tomasic. 2010. Word sense disambiguation via human computation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, page 60–63, New York, NY, USA. Association for Computing Machinery.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263.

Kathleen Tuite. 2014. Gwaps: Games with a problem. In *FDG*.

Noortje J. Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 397–403, Potsdam, Germany. Association for Computational Linguistics.

Luis Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Luis Von Ahn and Laura Dabbish. 2005. Esp: Labeling images with a computer game. In *AAAI spring symposium: Knowledge collection from volunteer contributors*, volume 2.

Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006a. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78.

Luis Von Ahn, Ruoran Liu, and Manuel Blum. 2006b. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 55–64.

Jérôme Waldispühl, Attila Szantner, Rob Knight, Sébastien Caisse, and Randy Pitchford. 2020. Leveling up citizen science. *Nature Biotechnology*, 38:1124–1126.

Mingzhu Wu, Nafise Sadat Moosavi, Dan Roth, and Iryna Gurevych. 2021. Coreference reasoning in machine reading comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5768–5781, Online. Association for Computational Linguistics.