

# WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles

Abbas Ghaddar, Philippe Langlais

RALI-DIRO

Montreal, Canada

abbas.ghaddar@umontreal.ca, felipe@iro.umontreal.ca

## Abstract

This paper presents WikiCoref, an English corpus annotated for anaphoric relations, where all documents are from the English version of Wikipedia. Our annotation scheme follows the one of OntoNotes with a few disparities. We annotated each markable with coreference type, mention type and the equivalent Freebase topic. Since most similar annotation efforts concentrate on very specific types of written text, mainly newswire, there is a lack of resources for otherwise over-used Wikipedia texts. The corpus described in this paper addresses this issue. We present a freely available resource we initially devised for improving coreference resolution algorithms dedicated to Wikipedia texts. Our corpus has no restriction on the topics of the documents being annotated, and documents of various sizes have been considered for annotation.

**Keywords:** Annotated Corpus, Coreference Resolution, Wikipedia

## 1. Introduction

In the last decade, coreference resolution has received an increasing interest from the NLP community, and became a standalone task in conferences and competitions due its role in applications such as Question Answering (QA), Information Extraction (IE), etc. This can be observed through, either the growth of coreference resolution systems varying from machine learning approaches e.g (Haghighi and Klein, 2009) to rule based systems e.g. (Lee et al., 2013), or the large-scale of annotated corpora comprising different text genres and languages.

Wikipedia<sup>1</sup> is a very large multilingual, domain-independent encyclopedic repository. The English version, as of July 2015, contains more than 4M articles, thus providing a large coverage of knowledge resources. Wikipedia articles are highly structured and follow strict guidelines and policies. Not only are articles formatted into sections and paragraphs, moreover volunteer contributors are expected to follow a number of rules<sup>2</sup> (specific grammars, vocabulary choice and other language specifications) that makes Wikipedia articles a text genre of its own.

Over the past few years, Wikipedia imposed itself on coreference resolution systems as a semantic knowledge source, owing to its highly structured organization and especially to a number of useful reference features such as redirects, out links, disambiguation pages, and categories. Although the boost in English annotated corpora tagged with anaphoric coreference relations and attributes, none of them involve Wikipedia articles as its main component.

This matter of fact motivated us to annotate Wikipedia documents for coreference, with the hope that it will foster research dedicated to this type of text. We introduce WikiCoref, an English corpus, constructed purely from Wikipedia articles, with the main objective to balance top-

ics and text size. This corpus has been annotated neatly by embedding state-of-the art tools (a coreference resolution system as well as a Wikipedia/FreeBase entity detector) that were used to assist manual annotation. This phase was then followed by a correction step to ensure fine quality. Our annotation scheme is mostly similar to the one followed within the OntoNotes project (Pradhan et al., 2007), yet with some minor differences.

Contrary to similar endeavours (see Section 2 for an overview), the project described here is small, both in terms of budget and corpus size. Still, one annotator managed to annotate 7955 mentions in 1785 coreference chains among 30 documents of various sizes, thanks to our semi-automatic named entity tracker approach. The quality of the annotation has been measured on a subset of three documents annotated by two annotators. The current corpus is in its first release, and will be upgraded in terms of size (more topics) in subsequent releases.

The remainder of paper is organized as follows. Section 2, discusses recent related works. We describe the annotation process in Section 3. In Section 4, we present our annotation scheme along with a detailed description of attributes assigned to each mention. We present in Section 5 the main statistics of our corpus. Annotation reliability is measured in Section 6, before ending the paper with conclusions and future works.

## 2. Related Work

In the last two decades, coreference resolution imposed itself on the natural language processing community as an independent task in a series of evaluation campaigns. This gave birth to various corpora designed in part to support training, adapting or evaluating of coreference resolution systems.

It began with the Message Understanding Conferences in which a number of comprehension tasks have been defined. Two resources have been designed within those

---

<sup>1</sup><https://www.wikipedia.org/>

<sup>2</sup>[https://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style)

tasks: the so-called MUC-6 and MUC-7 datasets created in 1995 and 1997 respectively (Hirshman and Chinchor, 1998). Those resources annotate named entities and coreferences on newswire articles.

A succeeding work is the Automatic Content Extraction (ACE) program monitoring tasks such as Entity Detection and Tracking (EDT). The so-called ACE-corpus has been released several times. The first release (Doddington et al., 2004) initially included named entities and coreference annotations for texts extracted from the TDT collection which contains newswire, newspaper and broadcast text genres. The last release extends the size of the corpus from 100k to 300k tokens (English part) and annotates other text genres (dialogues, weblogs and forums).

The OntoNotes project (Pradhan et al., 2007) is a collaborative annotation effort conducted by BBN Technologies and several universities, which aims is to provide a corpus annotated with syntax, propositional structure, named entities and word senses, as well as coreference resolution. The corpus reached its final release (5.0) in 2013, exceeding all previous resources with roughly 1.5 million of English words. It includes texts from five different text genres: broadcast conversation (200k), broadcast news (200k), magazine (120k), newswire (625k), and web data (300k). This corpus was for instance used within the CoNLL-2011 shared task (Pradhan et al., 2011) dedicated to entity and event coreference detection.

All those corpora are distributed by the Linguistic Data Consortium (LDC)<sup>3</sup>, and are largely used by researchers to develop and compare their systems. It is important to note that most of the annotated data originates from news articles. Furthermore, some studies (Hendrickx and Hoste, 2009; Nicolov et al., 2008) have demonstrated that a coreference resolution system trained on newswire data performs poorly when tested on other text genres. Thus, there is a crucial need for annotated material of more text genres and domains. This need has been partially fulfilled by some initiatives we describe hereafter.

Rodriguez et al. (2010) as part of the Live Memories project, present an Italian corpus annotated for anaphoric relations. The Corpus contains texts from the Italian Wikipedia and from blog sites with users comments. The selection of topics was restricted to historical, geographical, and cultural items, related to Trentino-Alto Adige/Sudtirolo, a region of North Italy. Poesio (2004) studies new text genres in the GNOME corpus. The corpus includes texts from three domains: Museum labels describing museum objects and artists that produced them, leaflets that provide information about patients medicine, and dialogues selected from the Sherlock corpus (Poesio et al., 2002).

Coreference resolution on biomedical texts took its place as an independent task in the BioNLP field; see for instance the Protein/Gene coreference task at BioNLP 2011 (Nguyen et al., 2011). Corpora supporting biomedical coreference tasks follow several annotation schemes and

domains. The MEDCo<sup>4</sup> corpus is composed of two text genres: abstracts and full papers. MEDSTRACT (Castano et al., 2002) consists of abstracts only, and DrugNerAr (Segura-Bedmar et al., 2010) annotates texts from the DrugBank corpus. The three aforementioned works follow the annotation scheme used in MUC-7 corpus, and restrict markables to a set of biomedical entity types. On the contrary, the CRAFT project (Cohen et al., 2010) adopts the OntoNotes guidelines and marks all possible mentions. The authors reported however a Krippendorff's alpha (Klaus, 1980) coefficient of only 61.9%.

Last, it is worth mentioning the (Schäfer et al., 2012) corpus gathering 266 scientific papers from the ACL anthology (NLP domain) and annotated with coreference information and mention type tags. In spite of partly garbled data (due information lost during the pdf conversion step) and low inter-annotator agreement, the corpus is considered a step forward in the coreference domain.

### 3. Methodology

In this section we describe how we selected the material to annotate in WikiCoref, the automatic preprocessing of the documents we conducted in order to facilitate the annotation task, as well as the annotation toolkit we used.

#### 3.1 Article Selection

We tried to build a balanced corpus in terms of article types and length, as well as in the number of out links they contain. We describe hereafter how we selected the articles to annotate according to each criterion.

A quick inspection of Wikipedia articles reveals that more than 35% of them are one paragraph long (that is, contain less than 100 words) and that only 11% of them contains 1000 words or more. We sampled articles of at least 200 words (too short documents are not very informative) paying attention to have a uniform sample of articles at size ranges [ $<1000$ ], [1000-2000], [2000-5000] and [ $>5000$ ].

We also paid attention to select articles based on the number of out links they contain. Out links encode a great part of the semantic knowledge embedded in an article. Thus, we paid attention to select evenly articles with high and low out link density. We further excluded articles that contain an overload of out links; normally those articles are indexes to other articles sharing the same topics, such as the article *List of President of the United States*.

In order to ensure that our corpus covers many topics of interest, we used the gazetteer generated by (Ratinov and Roth, 2009). It contains a collection of 16 (high precision low recall) lists of Wikipedia article titles that cover diverse topics, such as People, Organization, Human made Object, or Occupation. We selected our articles from all those lists, proportional to lists size.

<sup>3</sup><http://www ldc.upenn.edu/>

<sup>4</sup><http://nlp.i2r.a-star.edu.sg/medco.html>

### 3.2 Text Extraction

Although Wikipedia offers so-called Wikipedia dumps, parsing such files is rather tedious. Therefore we transformed the Wikipedia dump from its original XML format into the Berkeley database format compatible with WikipediaMiner (Milne and Witten, 2008). This system provides a neat Java API for accessing any piece of Wikipedia structure, including in and out links, categories, as well as a clean text (released of all Wikipedia markup).

Before preparing the data for annotation, we performed some slight manipulation of the data, such as removing the text of a bunch of specific sections (See also, Category, References, Further reading, Sources, Notes, and External links). Also, we removed section and paragraph titles. Last, we also removed ordered lists within an article as well as the preceding sentence. Those materials are of no interest in our context.

### 3.3 Markables Extraction

We used the Stanford CoreNLP toolkit (Manning et al., 2014), an extensible pipeline that provides core natural language analysis, to automatically extract candidate mentions along with high precision coreference chains, as explained shortly. The package includes the Dcoref multi-sieve system (Raghunathan et al., 2010; Lee et al., 2013), a deterministic coreference resolution rule-based system consisting of two phases: mention extraction and mention processing. Once the system identifies candidate mentions, it sends them, one by one, successively to ten sieves arranged from high to low precision in the hope that more accurate sieves will solve the case first. We took benefit of the system’s simplicity to extend it to the specificity of Wikipedia. We found these treatments described hereafter very useful in practice, notably for keeping track of coreferent mentions in large articles.

We first applied a number of pre-processing stages, benefiting from the wealth of knowledge and the high structure of Wikipedia articles. Each anchored text in Wikipedia links a human labelled span of text to one Wikipedia article. For each article we track the spans referring to it, to which we added the so-called redirects (typically misspellings and variations) found in the text, as well as the Freebase (Bollacker et al., 2008) aliases. When available in the Freebase

(a) On December 22, 2010, Obama signed [the Don’t Ask, Don’t Tell Repeal Act of 2010], fulfilling a key promise made in the 2008 presidential campaign...

(b) He signed into law [the Car Allowance Rebate System]<sub>x</sub>, known colloquially as [“Cash for Clunkers”]<sub>x</sub>, that temporarily boosted the economy.

Figure 1: Example of mentions detected (a) and linked (b) by our method.

structure we also collected attributes such as the type of the Wikipedia concept, as well as its gender and number attributes to be sent later to Stanford Dcoref.

All mentions that we detect this way allow us to extend Dcoref candidate list by mentions missed by the system (as in example -a- of Fig.1). Also, all mentions that refer to the same concept were linked into one coreference chain as in example -b-. This step greatly benefits the recall of the system as well as its precision, consequently our pre-processing method.

In addition, a mention detected by Dcoref is corrected when: a) a larger Wikipedia/Freebase mention exists, as in example -c- of Fig.2, b) a Wikipedia/Freebase mention shares some content words with a mention detected by Dcoref, as in example -d- of Fig.2.

(c) In December 2008, Time magazine named Obama as its [Person of <the Year><sub>Dcoref</sub>]<sub>Wiki/FB</sub> for his historic candidacy and election, which it described as “the steady march of seemingly impossible accomplishments”.

(d) Obama also introduced Deceptive Practices and Voter Intimidation Prevention Act, a bill to criminalize deceptive practices in federal elections, and [the Iraq War De-Escalation Act of <2007><sub>Wiki/FB</sub>, neither of which was signed into law><sub>Dcoref</sub>.

Figure 2: Examples of contradictions between Dcoref mentions (marked by angular brackets) and our method (marked by squared brackets)

Second, we applied some post-treatments on the output of the Dcoref system. First, we removed coreference links between mentions whenever it has been detected by a sieve other than: Exact Match (second sieve which links two mentions if they have the same string span including modifiers and determiners), Precise Constructs (forth sieve which recognizes two mentions are coreferential if one of the following relation exists between them: Appositive, Predicate nominative, Role appositive, Acronym, Demonym). Both sieves score over 95% in precision according to (Raghunathan et al., 2010). We do so to prevent as much as possible noisy mentions in the pre-annotation phase.

Overall, we corrected roughly 15% of the mentions detected by Dcoref, we added and linked over 2000 mentions for a total of 4318 ones, 3871 of which were found in the final annotated data.

### 3.4 Annotation Tool and Format

Manual annotation is performed using MMAX2 (Müller and Strube, 2006), which supports stand-off format. The toolkit allows multi-coding layers annotation at the same time and the graphical interface (Figure 3) introduces a multiple

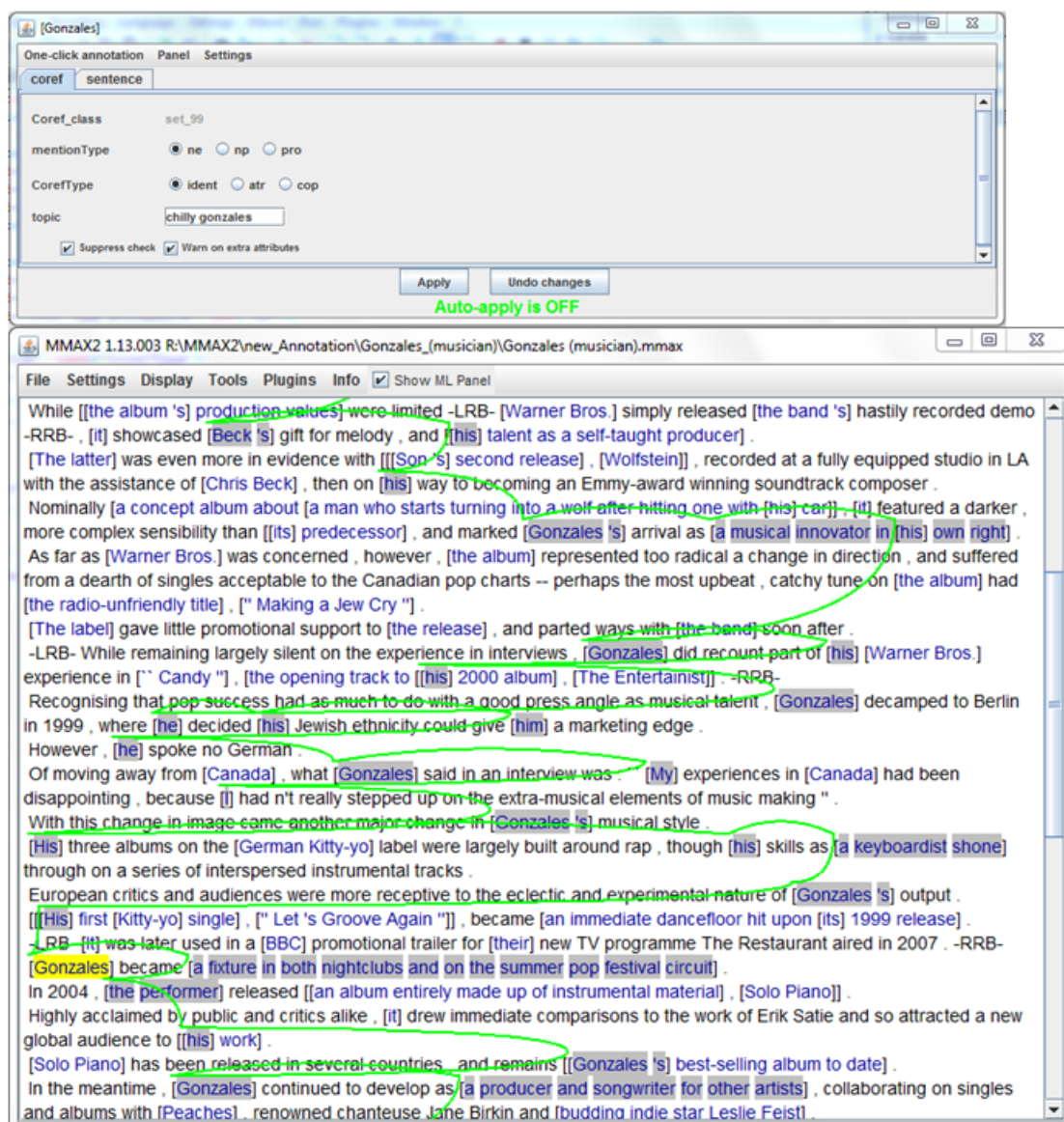


Figure 3: Annotation of WikiCoref in MMAX2 tool

pointer view in order to track coreference chain membership. Automatic annotations were transformed from Stanford XML format to the MMAX2 format previously to human annotation. The WikiCoref corpus is distributed in the MMAX2 stand-off format.

## 4. Annotation Scheme

In general, the annotation scheme in WikiCoref mainly follows the OntoNotes scheme (Pradhan et al., 2007). In particular, only noun phrases are eligible to be mentions and only non-singleton coreference sets are kept in the version distributed. Each annotated mention is tagged by a set of attributes: mention type (Section 4.1), coreference type (Section 4.2) and the equivalent Freebase topic when available (Section 4.3). In Section 4.4, we introduce a few modifications we made to the OntoNotes guidelines in order to reduce ambiguity, consequently optimize our inter-annotator agreement.

### 4.1 Mention Type

#### 4.1.1 Named entity (NE)

NEs can be proper names, NPs or abbreviations referring to an object in the real world. Typically, a named entity may be a person, an organization, an event, a facility, a geopolitical entity, etc. Our annotation is not tied to a limited set of named entities.

NEs are considered to be atomic, as a result, we omit the sub-mention *Montreal* in the full mention *University of Montreal*, as well as units of measures and expressions referring to money if they occur within a numerical entity, e.g. Celsius and Euro signs in the mentions *30 C °* and *1000 €* are not marked independently. If the mention span is a named entity and it is preceded by the definite article *'the'* (who refers to the entity itself), we add the latter to the span and the mention type is always NE. For instance, in *The United States* the whole span is marked as a NE.

### 4.1.2 Noun Phrase (NP)

Noun phrase (group of words headed by a noun, or pronouns) mentions are marked as NP when they are not classified as Named entity. The NP tag gathers three noun phrase type. **Definite Noun Phrase**, designates noun phrases which have a *definite description* usually beginning with the definite article *the*. **Indefinite Noun Phrase**, are noun phrases that have an *indefinite description*, mostly phrases that are identified by the presence of the indefinite articles *a* and *an* or the absence of determiners. **Conjunction Phrase**, that is, at least two NPs connected by a coordinating or correlative conjunction (e.g. *the man and his wife*), for this type of noun phrase we don't annotate discontinuous markables. However, unlike named entities we annotate mentions embedded within NP mentions whatever the type of the mention is. For example, we mark the pronoun *his* in the NP mention *his father*, and *Obama* in *the Obama family*.

### 4.1.3 Pronominal (PRO)

Mentions tagged PRO may be one of the following subtypes: personal, possessive, reflexive, and demonstrative pronouns.

## 4.2 Coreference Type

MUC and ACE schemes treat identical and attributive mentions as coreferential, contrary to the OntoNotes scheme which differentiates between these two because they play different roles. In addition, OntoNotes omits attributes signaled by copular structures. To be as much as possible faithful to those annotation schemes, we tag as identical (IDENT) all referential mentions; as attributive (ATR) all mentions in appositive (e.g. example -e- of Fig. 4), parenthetical (example -f-) or role appositive (example -g-) relation; and lastly Copular (COP) attributive mentions in copular structures (example -h-). We added the latest because it offers useful information for coreference systems.

(e) [Jefferson Davis] <sub>ATR</sub> , [President of the Confederate States of America] <sub>ATR</sub>
(f) [The Prime Minister's Office] <sub>ATR</sub> ([PMO] <sub>ATR</sub> ) .
(g) [The Conservative lawyer] <sub>ATR</sub> [John P. Chipman] <sub>ATR</sub>
(h) Borden is [the chancellor of Queen's University] <sub>COP</sub>

Figure 4: Example of Attributive mentions

### 4.3 Freebase Attribute

At the end of the annotation process we assign for each coreference chain the corresponding Freebase entity (knowing that the equivalent Wikipedia link is already included in the Freebase dataset). We think that this attribute

will facilitate the extraction of features relevant to coreference resolution tasks, such as gender, number, animacy, etc. It also makes the corpus usable in wikification tasks.

## 4.4 Scheme Modifications

As mentioned before, our annotation scheme follows OntoNotes guidelines with slight adjustments. Besides marking predicate nominative attributes, we made two modifications to the OntoNotes guidelines that are described hereafter.

### 4.4.1 Maximal Extent

In our annotation, we identify the maximal extent of the mention, thus including all modifiers of the mention: either pre-modifiers like determiners or adjectives modifying the mention, or post-modifiers like prepositional phrases (e.g. *The federal Cabinet also appoints justices to [superior courts in the provincial and territorial jurisdictions]*), relative clauses phrases (e.g. *[The Longueuil International Percussion Festival which features 500 musicians], takes place...*).

Otherwise said, we only annotate the full mentions contrary to those examples extracted from OntoNotes where sub-mentions are also annotated:

- [ [Zsa Zsa] <sub>X</sub>, who slap a security guard ] <sub>X</sub>
- [ [a colorful array] <sub>X</sub> of magazines ] <sub>X</sub>

### 4.4.2 Verbs

Our annotation scheme does not support verbs or NP referring to them inclusively.

## 5. Corpus Description

The first release of the WikiCoref corpus consists of 30 documents, comprising 59,652 tokens spread over 2,229 sentences. Document size varies from 209 to 9,869 tokens; for an average of approximately 2000 tokens. Table 1 summarizes the main characteristics of a number of existing coreference-annotated corpora. Our corpus is the smallest in terms of the number of documents but is comparable

Corpus	Size	#Doc	#Doc/Size
ACE-2007 (English)	300k	599	500
(Schäfer et al., 2012)	1.33M	226	4986
LiveMemories (Italian)	150k	210	714
MUC-6	30k	60	500
MUC-7	25k	50	500
OntoNotes 1.0	300k	597	502
WikiCoref	60k	30	2000

Table 1: Main characteristics of WikiCoref compared to existing coreference-annotated corpora

in token size with some other initiatives, which we believe makes it already a useful resource.

The distribution of coreference and mentions types is presented in Table 2. We observe the dominance of NE mentions 45% over NP ones 40%, an unusual distribution we believe to be specific to Wikipedia. As a matter of fact, concepts in this resource (e.g. *Barack Obama*) are often referred by their name or a variant (e.g. *Obama*) instead of an NP (e.g. *the president*). In (Schäfer et al., 2012) the authors observe for instance that only 22.1% of mentions are named entities in their corpus of scientific articles.

Mention Type	Coreference Type			Total
	IDENT	ATR	COP	
NE	3279	258	20	3557
NP	2489	388	296	3173
PRO	1225	-	-	1225
<b>Total</b>	6993	646	316	7955

Table 2: Frequency of mention and coreference types in WikiCoref

We annotated 7286 identical and copular attributive mentions that are spread into 1469 coreference chains, giving an average chain length of 5. The distribution of chain length is provided in Figure 5. Also, WikiCoref contains 646 attributive mentions distributed over 330 attributive chains.

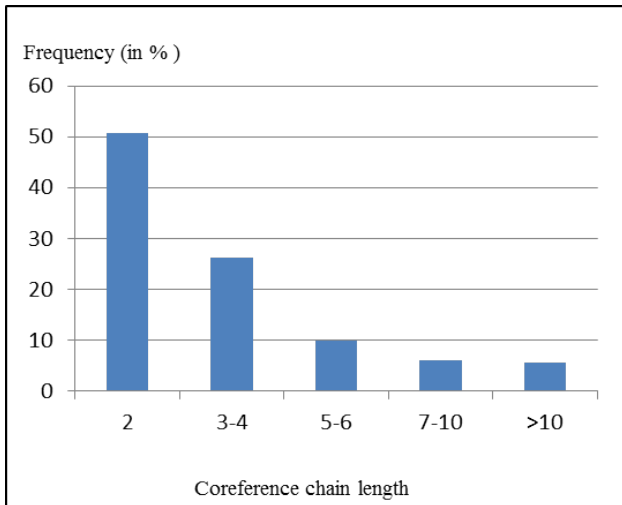


Figure 5: Distribution of the coreference chains length

We observe that half of the chains have only two mentions, and that roughly 5.7% of the chains gather 10 mentions or more. In particular, the concept described in each Wikipedia article has an average of 68 mentions per document, which represents 25% of the WikiCoref mentions.

## 6. Inter-Annotator Agreement

Coreference annotation is a very subtle task which involves a deep comprehension of the text being annotated, and a rather good sense of linguistic skills for smartly applying

the recommendations in annotation guidelines. Most of the material currently available has been annotated by the first author only. In an attempt to measure the quality of the annotations produced, we asked another annotator to annotate 3 documents already treated by the first annotator. The subset of 5520 tokens represents 10% of the full corpus in terms of tokens. The second annotator had access to the OntoNotes guideline (Pradhan et al., 2007) as well as to a bunch of selected examples we extracted from the OntoNotes corpus.

On the task of annotating mention identification, we measured a Kappa coefficient (Carletta, 1996) of 0.78, which is slightly close to the well accepted threshold of 80%, but it falls in the range of other endeavors and it roughly indicates that both subjects often agreed.

We also measured a MUC F1 score (Vilain et al., 1995) of 83.3%. We computed this metric by considering one annotation as ‘Gold’ and the other annotation as ‘Response’, the same way coreference system responses are evaluated against Key annotations. In comparison to (Schäfer et al., 2012) who reported a MUC of 49.5, it’s rather encouraging for a first release. This sort of indicates that the overall agreement in our corpus is acceptable.

## 7. Conclusions

We presented WikiCoref, a coreference-annotated corpus made merely from English Wikipedia articles. Documents were selected carefully to cover various stylistic articles.

Each mention is tagged with syntactic and coreference attributes along with its equivalent Freebase topic, thus making the corpus eligible to both training and testing coreference systems; our initial motivation for designing this resource. The annotation scheme followed in this project is an extension of the OntoNotes scheme.

To measure inter-annotators agreement of our corpus, we computed the Kappa and MUC scores, both suggesting a fair amount of agreement in annotation. The first release of WikiCoref can be freely downloaded at <http://rali.iro.umontreal.ca/rali/?q=en/wikicoref>.

We hope that the NLP community will find it useful and plan to release further versions covering more topics.

## 8. Acknowledgements

This work has been funded by Nuance Foundation. We are grateful to Fabrizio Gotti who kindly took part in the annotation process, and assisted us to refine our annotation scheme. Also, we would like to thank the reviewers for helpful comments.

## 9. References

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph

- database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.
- Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.
- Castano, J., Zhang, J., and Pustejovsky, J. (2002). Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution*.
- Cohen, K. B., Lanfranchi, A., Corvey, W., Baumgartner Jr, W. A., Roeder, C., Ogren, P. V., Palmer, M., and Hunter, L. (2010). Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *Proceedings of BioTxtM 2010*, pages 37–41.
- Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ACE) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1.
- Haghighi, A. and Klein, D. (2009). Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161.
- Hendrickx, I. and Hoste, V. (2009). Coreference resolution on blogs and commented news. In *Anaphora Processing and Applications*, pages 43–53.
- Hirshman, L. and Chinchor, N. (1998). MUC-7 coreference task definition. version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.
- Klaus, K. (1980). *Content analysis: An introduction to its methodology*.
- Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.
- Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.
- Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, pages 197–214.
- Nguyen, N., Kim, J. D., and Tsujii, J. (2011). Overview of bionlp 2011 protein coreference shared task. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 74–82.
- Nicolov, N., Salvetti, F., and Ivanova, S. (2008). Sentiment analysis: Does coreference matter. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 37.
- Poesio, M., Di Eugenio, B., and Keohane, G. (2002). Discourse structure and anaphora: An empirical study. *Rapport Technique NLE Technical Note TN-02-02. University of Essex*.
- Poesio, M. (2004). Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79.
- Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J., and Micciulla, L. (2007). Unrestricted coreference: Identifying entities and events in OntoNotes. In *null*, pages 446–453.
- Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.
- Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.
- Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155.
- Rodriguez, K. J., Delogu, F., Versley, Y., Stemle, E. W., and Poesio, M. (2010). Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proceedings of LREC*, pages 157–163.
- Schäfer, U., Spurk, C., and Steffen, J. (2012). A fully coreference-annotated corpus of scholarly papers from the acl anthology. In *Proceedings of COLING 2012*, pages 1059–1070.
- Segura-Bedmar, I., Crespo, M., de Pablo-Sánchez, C., and Martínez, P. (2010). Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC bioinformatics*, 11:1–11.
- Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52.