Université de Montréal

Coreference Resolution with and for Wikipedia

par
Abbas Ghaddar

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en computer science

Juin, 2016

RÉSUMÉ

Wikipédia est une ressource embarquée dans de nombreuses applications du traitement des langues naturelles. Pourtant, aucune étude à notre connaissance n'a tenté de mesurer la qualité de résolution de coréférence dans les textes de Wikipédia, une étape préliminaire à la compréhension de textes. La première partie de ce mémoire consiste à construire un corpus de coréférence en anglais, construit uniquement à partir des articles de Wikipédia. Les mentions sont étiquetées par des informations syntaxiques et sémantiques, avec lorsque cela est possible un lien vers les entités FreeBase équivalentes. Le but est de créer un corpus équilibré regroupant des articles de divers sujets et tailles. Notre schéma d'annotation est similaire à celui suivi dans le projet OntoNotes. Dans la deuxième partie, nous allons mesurer la qualité des systèmes de détection de coréférence à l'état de l'art sur une tâche simple consistant à mesurer les mentions du concept décrit dans une page Wikipédia (p. ex : les mentions du président Obama dans la page Wikipédia dédiée à cette personne). Nous tenterons d'améliorer ces performances en faisant usage le plus possible des informations disponibles dans Wikipédia (catégories, redirects, infoboxes, etc.) et Freebase (information du genre, du nombre, type de relations avec autres entités, etc.).

Mots cles: Résolution de Coréférences, Création du corpus, Wikipedia

ABSTRACT

Wikipedia is a resource of choice exploited in many NLP applications, yet we are not aware of recent attempts to adapt coreference resolution to this resource, a preliminary step to understand Wikipedia texts. The first part of this master thesis is to build an English coreference corpus, where all documents are from the English version of Wikipedia. We annotated each markable with coreference type, mention type and the equivalent Freebase topic. Our corpus has no restriction on the topics of the documents being annotated, and documents of various sizes have been considered for annotation. Our annotation scheme follows the one of OntoNotes with a few disparities. In part two, we propose a testbed for evaluating coreference systems in a simple task of measuring the particulars of the concept described in a Wikipedia page (eg. The statements of President Obama the Wikipedia page dedicated to that person). We show that by exploiting the Wikipedia markup (categories, redirects, infoboxes, etc.) of a document, as well as links to external knowledge bases such as Freebase (information of the type, number, type of relationship with other entities, etc.), we can acquire useful information on entities that helps to classify mentions as coreferent or not.

Keywords: Coreference Resolution, Corpus Creation, Wikipedia.

CONTENTS

RÉSUMÉ . . . . . **ii**

ABSTRACT . . . . . **iii**

CONTENTS . . . . . **iv**

LIST OF TABLES . . . . . **vii**

LIST OF FIGURES . . . . . **ix**

ACKNOWLEDGMENTS . . . . . **xi**

CHAPTER 1: INTRODUCTION **1**

1.1 Introduction to Coreference resolution 1

1.2 Structure of the master thesis 3

1.3 Summary of Contributions 3

CHAPTER 2: RELATED WORK **4**

2.1 Coreference Annotated Corpora 4

2.2 State of the Art of Coreference Resolution Systems 6

2.3 Coreference Resolution Features 8

2.4 Evaluation Metrics 10

2.4.1 MUC 10

2.4.2 B^3 11

2.4.3 CEAF 11

2.4.4 BLANC 13

2.4.5 CoNLL score and state-of-the-art Systems 14

2.4.6 Wikipedia and Freebase 16

CHAPTER 3: WIKICOREF: AN ENGLISH COREFERENCE-ANNOTATED CORPUS OF WIKIPEDIA ARTICLES 18

3.1 Introduction 18
3.2 Methodology 19
3.2.1 Article Selection 19
3.2.2 Text Extraction 21
3.2.3 Markables Extraction 21
3.2.4 Annotation Tool and Format 24
3.3 Annotation Scheme 25
3.3.1 Mention Type 26
3.3.2 Coreference Type 28
3.3.3 Freebase Attribute 29
3.3.4 Scheme Modifications 29
3.4 Corpus Description 30
3.5 Inter-Annotator Agreement 33
3.6 Conclusions 33

CHAPTER 4: WIKIPEDIA MAIN CONCEPT DETECTOR 35

4.1 Introduction 35
4.2 Baselines 37
4.3 Approach 39
4.3.1 Preprocessing 39
4.3.2 Feature Extraction 40
4.4 Dataset 44
4.5 Experiments 45
4.5.1 Data Preparation 45
4.5.2 Classifier 46
4.5.3 Main Concept Resolution Performance 48
4.5.4 Coreference Resolution Performance 51
4.6 Conclusion 52

BIBLIOGRAPHY **54**

LIST OF TABLES

2.I Summary of the main coreference-annotated corpora 6

2.II The BLANC confusion matrix, the values of example of Figure 2.2 are placed between parenthesizes. 15

2.III Formula to calculate BLANC: precision recall and F1 score . . . 15

2.IV Performance of the top five systems in the CoNLL-2011 shared task 15

2.V Performance of current state-of-the-art systems on CoNLL 2012 English test set, including in order: [5]; [35]; [11]; [73] ; [74] . . 16

3.I Main characteristics of WikiCoref compared to existing coreference-annotated corpora 30

3.II Frequency of mention and coreference types in WikiCoref 31

4.I The eleven feature encoding string similarity (10 row) and semantic similarity (row number 11). Columns two and three contain possible values of strings representing the MC (title or alias...) and a mention (mention span or head...) respectively. The last row shows the WordNet similarity between MC and mention strings. . 42

4.II The non-pronominal mention *main* features family 43

4.III CoNLL F1 score of recent state-of-the-art systems on the WikiCoref dataset, and the 2012 OntoNotes test data for predicted mentions. 44

4.IV Configuration of the SVM classifier for both pronominal and non pronominal models 46

4.V Performance of the baselines on the task of identifying all MC coreferent mentions. 47

4.VI Performance of our approach on the pronominal mentions, as a function of the features. 48

4.VII Performance of our approach on the non-pronominal mentions, as a function of the features. 49

4.VIII Performance of `Dcoref++` on WikiCoref compared to state of the art systems, including in order: [31]; [19] - Final; [20] - Joint; [35] - Ranking:Latent; [11] - Statistical mode with clustering. 51

LIST OF FIGURES

1.1 Sentences extracted from the English portion of the ACE-2004 corpus 1

2.1 Example on calculating B^3 metric scores 12
2.2 Example of key (gold) and response (System) coreference chains . 14
2.3 Excerpt from the Wikipedia article *Barack Obama* 16
2.4 Excerpt of the Freebase page of *Barack Obama* 17

3.1 Distribution of Wikipedia article depending on word count 20
3.2 Distribution of Wikipedia article depending on link density 20
3.3 Example of mentions detected by our method. 22
3.4 Example of mentions linked by our method. 22
3.5 Examples of contradictions between Dcoref mentions (marked by angular brackets) and our method (marked by squared brackets) . 23
3.6 Examples of contradictions between Dcoref mentions (marked by angular brackets) and our method (marked by squared brackets) . 24
3.7 Annotation of WikiCoref in MMAX2 tool 25
3.8 The XML format of the MMAX2 tool 26
3.9 Example of Attributive and Copular mentions 28
3.10 Example of Metonymy and Acronym mentions 29
3.11 Distribution of the coreference chains length 31
3.12 Distribution of distances between two successive mentions in the same coreference chain 32

4.1 Output of a CR system applied on the Wikipedia article *Barack Obama* 37
4.2 Representation of a mention. 40

4.3 Representation of a Wikipedia concept. The source from which the information is extracted is indicated in parentheses: (W)ikipedia, (F)reebase. 41
4.4 Examples of mentions (underlined) associated with the MC. An asterisk indicates wrong decisions. 50

ACKNOWLEDGMENTS

I am deeply grateful to Professor Philippe Langlais who is a fantastic supervisor, the last year has been intellectually stimulating, rewarding and fun. He has gently shepherded my research down interesting paths. I hope that I have managed to absorb just some of his dedication and taste in research, he is a true privilege.

I have been very lucky to meet and interact with the extraordinarily skillful Fabrizio Gotti who kindly help me to debug code when I got stuck on some computer problem. Also, he took part in the annotation process and assisted me to refine our annotation scheme.

Many thanks also to the members of RALI-lab that I have been fortunate enough to be surrounded by such a group of friends and colleagues.

I would like to thank my dearest parents, grandparent, aunt and uncle for being unwavering in their support.

CHAPTER 1

INTRODUCTION

1.1 Introduction to Coreference resolution

Coreference Resolution (CR) is the task of identifying all textual expressions that refer to the same entity. Entities are objects in the real or hypothetical world. The textual reference to an entity in the document is called *mention*. It can be a pronominal phrase (e.g. he), a nominal phrase (e.g. the performer) or a named entity (e.g. Chilly Gonzales). Two or more mentions are coreferring to each other if all of them resolve to a unique entity. The set of coreferential mentions form a *chain*. Consequently, mentions that are not part of any coreferential relation are called *singletons*. Consider the following example extracted from the 2004 ACE [18] dataset:

> [Eyewitnesses]$_{m1}$ reported that [Palestinians]$_{m2}$ demonstrated today Sunday in [the West Bank]$_{m3}$ against [the [Sharm el-Sheikh]$_{m4}$ summit to be held in [Egypt]$_{m6}$]$_{m5}$. In [Ramallah]$_{m7}$, [around 500 people]$_{m8}$ took to [[the town]$_{m9}$'s streets]$_{m10}$ chanting [slogans]$_{m11}$ denouncing [the summit]$_{m12}$ and calling on [Palestinian leader Yasser Arafat]$_{m13}$ not to take part in [it]$_{m14}$.

Figure 1.1 – Sentences extracted from the English portion of the ACE-2004 corpus

A Typical CR system will output {m5, m12, m14} and {m7, m9} as two coreference chains and the rest as singletons. The three mentions in the first chain are referent to *"the summit held in Egypt"*, while the second chain is equivalent to *"the town of Ramallah"*. Human knowledge gives people the ability to easily infer such relations, but it turns out to be extremely challenging for automated systems. However, coreference resolution requires a combination of different kinds of linguistic knowledge, discourse processing, and semantic knowledge. Sometimes, CR is confused with the similar task of anaphora resolution. The goal of the latter is to find a referential relation (anaphora) between one

mention called *anaphor* and one of its *antecedent* mentions, where the antecedent is required for the interpretation of the anaphor. While CR aims to establish which noun phrases (NPs) in the text points to the same discourse entity. Thus, not all anaphoric cases can be treated as coreferential and vice versa. For example the bound anaphora relation between *dog* and *its* in the sentence *Every dog has its day*, is not considered as coreferential.

To its importance, CR is a prerequisite for various NLP tasks including information extraction [75], information retrieval [52], question answering [40], machine translation [29] and text summarization [4]. For example, in Open Information Extraction (OIE) [79], one acquires subject-predicate-object relations, many of which (e.g., <the foundation stone, was laid by, the Queen‘s daughter>) being useless because the subject or the object contains material coreferring to other mentions in the text being mined.

The first automatic coreference resolution systems handled the task with hand-crafted rules. In the 1970s, the problematic is limited to the resolution of pronominal anaphora, the first proposed algorithm [26] mainly explore the syntactic parse tree of the sentences. It making use of constraints and preferences on pronouns depending on its position in the tree. The latter works succeeded by a set of endeavours [1, 7, 30, 65] based on heuristics, thus only in the mid-1990 became available coreference-annotated corpora that eased to solve the problem with machine learning approaches.

The availability of large datasets annotated with coreference information change the focusing on supervised learning approaches, which leads to reformulate the identification of a coreference chain as a classification or clustering problem. It also fostered the elaboration of several evaluation metrics in order to evaluate the performance of a well-designed system.

While Wikipedia is ubiquitous in the NLP community, we are not aware of much works that involve Wikipedia articles in a coreference corpus or conducted to adapt CR to Wikipedia text genre.

1.2 Structure of the master thesis

This thesis addresses the problem of Coreference resolution in Wikipedia. In chapter 2 we review coreference resolution components: divers corpora annotated with coreference information used for training and testing; important approaches that influenced the domain; the most commonly used features in previous literature; and evaluation metrics adopted by the community. Chapter 3 is dedicated to the coreference-annotated corpus of Wikipedia article I created. Chapter 4 describe the work on the Wikipedia main concept mention detector.

1.3 Summary of Contributions

Chapter 3 and 4 of this thesis have been published in:

1. *Abbas Ghaddar and Phillippe Langlais. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), May 2016.*

2. *Abbas Ghaddar and Phillippe Langlais. Coreference in wikipedia: Main concept resolution. In Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL 2016), Berlin, Germany, August 2016.*

We elaborated a number of resources that the community can use:

1. Wikicoref: An english coreference-annotated corpus of wikipedia articles, avaivalble at
   `http://rali.iro.umontreal.ca/rali/?q=en/wikicoref`

2. A full English Wikipedia dump of April 2013, where all mentions corefering to the main concept are automatically extracted using the classifier described in Chapetr 4, along with information we extracted from Wikipedia and Freebase. The resource is available at
   `http://rali.iro.umontreal.ca/rali/en/wikipedia-main-concept`

CHAPTER 2

RELATED WORK

2.1 Coreference Annotated Corpora

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 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 tasks: the so-called MUC-6 and MUC-7 datasets created in 1995 and 1997 respectively [21, 25]. Those resources annotate named entities and coreferences on newswire articles. The MUC coreference annotation scheme consider NPs that refer to the same entity as *markables*. It support a wide coverage of coreference relations under the identity tag, such as predicative NPs and bound anaphors.

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 [18] 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 ACE corpus follows a well-defined annotation scheme, which distinguishes various relational phenomenon and assign to each mention a class attribute: Negatively Quantified, Attributive, Specific Referential, Generic Referential or Underspecified Referential [17]. Also, ACE restricts the type of entities to be annotated to seven: person, organization, geo-political, location, facility, vehicle, and weapon.

The OntoNotes project [57] 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 project extends the task definition to include verbs and events, also it tags mentions with two types of coreference: Identical (IDENT), and Appositive (APPOS), this will be detailed in the next chapter. 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 [54] dedicated to entity and event coreference detection.

All those corpora are distributed by the Linguistic Data Consortium (LDC) [1], 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 [24, 48] 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 different text genres and domains. This need has been partially fulfilled by some initiatives we describe hereafter.

The Live Memories project [66] introduces 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 AdigeSudtirol, a region of North Italy. Poesio et al.,[50] 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 [51].

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 [47]. Corpora supporting biomedical coreference tasks follow several annotation schemes and domains. The MEDCo [2] corpus is composed of two text genres: abstracts

1. http://www.ldc.upenn.edu/
2. http://nlp.i2r.a-star.edu.sg/medco.html

and full papers. MEDSTRACT [9] consists of abstracts only, and DrugNerAr [68] 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 [12] adopts the OntoNotes guidelines and marks all possible mentions. The authors reported however a Krippendorff's alpha [28] coefficient of only 61.9%.

Last, it is worth mentioning the corpus of [67] 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 to information lost during the pdf conversion step) and low inter-annotator agreement, the corpus is considered a step forward in the coreference domain. Table 2.I summarizes the aforementioned corpora that have been annotated with coreference information.

Year	**Corpus**	**Domain**	**Size**
1996	MUC-6	News	30k
1997	MUC-7	News	25k
2004	GNOME	Museum labels, leaflets and dialogues	50k
2005	ACE	News and weblogs	350k
2007	ACE	News, weblogs, dialogues and forums	300k
2007	OntoNotes 1.0	News	300k
2008	OntoNotes 2.0	News	500k
2010	LiveMemories (Italian)	News, blogs, Wikipedia, dialogues	150k
2008	[67]	NLP scientific paper	1.33M
2013	OntoNotes 5.0	conversation, magazine, newswire, and web data	1.5M

Table 2.I – Summary of the main coreference-annotated corpora

2.2 State of the Art of Coreference Resolution Systems

Different types of approaches differ as to how to formulate the task entrusted to learning algorithms, including:

Pairwise models [69] : are based on a binary classification comparing an anaphora to potential antecedents located in previous sentences. Specifically, the examples provided to the model are mentions pairs (anaphora and a potential antecedent) for which the objective of the model is to determine whether the pair is coreferent or not. In a second phase, the model determines which mention pairs can be classified as coreferent, and the real antecedent of an anaphora from all its antecedent coreferent mentions. Those models are widely used and various systems have implemented them, such as [3, 44, 45] to cite a few.

Twin-candidate models [77] As in pairwise models, the problem is considered as a classification task, but whose instances are composed of three elements (x, y_i, y_j) where x is an anaphora and y_i, y_j are two antecedents candidates (where y_i is the closest to x in terms of distance). The purpose of the model is to establish a criteria for comparing the two antecedents for this anaphora, and rank y_i as FIRST if it's the best antecedent or as SECOND if y_j is the best antecedent. This classification alternative is interesting because it no longer considers the resolution of the coreference as the addition of independent anaphoric resolutions (mention pairs), but considers the "competitive" aspect of the various possible antecedents for anaphora.

Mention-ranking models : the model was initially proposed by [15], it doesn't aim to classify pairs of mentions but to classify all possible antecedents for a given anaphora in an iterative process. The process successively compares an anaphora with two potential antecedents. At each iteration, the best candidate is stored and then form a new pair of candidates with the "winner" and the new candidate. The iteration stops when no more possible candidate is left. An alternative to this method is to simultaneously compare all possible histories for a given anaphora. The model was implemented in [19, 59] to cite a few.

Entity-mention models [78] : They determine the probability of a mention referring to an entity or to an entity cluster using a vector of coreference feature level and cluster (i.e. a candidate is compared to a single antecedent or a cluster con-

taining all references to the same entity). The model was implemented in [33, 78]

Multi-sieve models [58] : Once the model identifies candidate mentions, it sends a mention and its antecedent to sieves arranged from high to low precision, in the hope that more accurate sieves will merge the mention pair under a single cluster. The model was implemented by a rule-based system [31] as well as in machine learning system [62].

2.3 Coreference Resolution Features

Most CR systems focus on syntactic and semantic characteristics of mention to decide which mentions should be clustered together. Given a mention m_i and an antecedent mention m_j, we list the most common used features that enable a CR system to capture coreference between mentions. We classify the features into four categories: **String Similarity** ([45, 58, 69]); **Semantic Similarity** ([14, 31, 44]); **Relative Location** ([3, 22, 43]); and **External Knowledge** ([22, 23, 43, 53, 62]).

String Similarity: This family of features indicate that m_i and m_j are coreferent by looking to if their strings share some properties, such as:

- String match (without determiners);
- m_i and m_j are pronominal/proper names/non-pronominal and the same string;
- m_i and m_j are proper names/non-pronominal and one is a substring of the other;
- The words of m_i and m_j intersect;
- Minimum edit distance between m_i and m_j string;
- Head match;
- m_i and m_j are part of a quoted string;
- m_i and m_j have the same maximal NP projection;
- One mention is an acronym of the other;
- Number of different capitalized words in two mentions;
- Modifiers match;
- The pronominal modifiers of one mention are a subset of those of the other;

- Aligned modifiers relation.

Semantic Similarity: Captures the semantic relation between two mentions by enforcing agreement constraints between them.

- Number agreement;
- Gender agreement;
- Mention type agreement;
- Animacy agreement;
- One mention is an alias of the other;
- Semantic class agreement;
- m_i and m_j are not proper names but contain mismatching proper names;
- Saliency;
- Semantic role.

Relative Location: Encode the distance between the two mentions on different layers.

- m_j is an appositive of m_i;
- m_j is a nominal predicate of m_i;
- Parse tree path from m_j to m_i;
- Word distance between m_j and m_i;
- Sentence distance between m_j and m_i;
- Mention distance between m_j and m_i;
- Paragraph distance between m_j and m_i.

External Knowledge: Try to link mentions to external knowledge in order to extract attributes that will be used during inference process.

- m_i and m_j have ancestor-descendent relationship in WordNet;
- One mention is a synonym/antonym/hypernym of the other in WordNet;
- WordNet similarity score for all synset pairs of m_i and m_j;
- The first paragraph of the Wikipedia page titled m_i contains m_j (or vice versa);
- The Wikipedia page titled m_i contains an hyperlink to the Wikipedia page

titled m_j (or vice versa);

- The Wikipedia page of m_i and the Wikipedia page of m_j have a common Wikipedia category..

2.4 Evaluation Metrics

In evaluation, we need to compare the true set of entities (KEY, produced by human expert) with the predicted set of entities (SYS, produced by the system). The task of coreference resolution is traditionally evaluated according to four metrics widely used in the literature. Each metric is computed in terms of recall (R), a measure of completeness, and precision (P), a measure of exactness and the F-score corresponds to the harmonic mean: F-score = 2 · P · R / (P + R).

2.4.1 MUC

The name of the MUC metric [72] is derived from the evaluation campaign *Message Understanding Conference*. This is the first and widelyused metric for scoring CR systems. The MUC score is calculated by identifying the minimum number of link modifications required to make the set of mentions identified by the system as coreferring perfectly align to the gold-standard set (called Key). That is, the total number of mentions minus the number of entities, otherwise said, it is the number of common links in key and system set. Let S_i designate a coreference chain returned by a system and G_i is a chain in the key reference. Consequently, $p(S_i)$ and $p(G_i)$ are chains of S_i and G_i relative to the system response and key respectively. That is, $p(S_i)$ is a chain and S_i is a mention in that chain. The following are respectively the formula for Precision, Recall and F1:

$$Precision = \frac{\sum(|G_i| - |p(G_i)|)}{\sum(|G_i| - 1)} \tag{2.1}$$

$$Recall = \frac{\sum(|S_i| - |p(S_i)|)}{\sum(|S_i| - 1)} \tag{2.2}$$

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{2.3}$$

For example, a key and a response are provided as below: key = {a,b,c,d} and response = {a,b},{c,d}. The MUC precision, recall and F-score for the example are calculated as:

$Precision = \frac{4-2}{4-1} = 0.66$
$Recall = \frac{(2-1)+(2-1)}{(2-1)+(2-1)} = 1.0$
$F1 = \frac{2 \cdot 2/3 \cdot 1}{2/3+1} = 0.79$

2.4.2 B^3

Bagga and Baldwin [2] present their B-CUBED evaluation algorithm to deal with three issues of the MUC-metric: only gain points for links, all errors are considered equal, and singleton mentions are not represented. Instead of looking at the links, B-CUBED metric measures the accuracy of coreference resolution based on individual mentions. Let R_{mi} be the response chain of mention m_i and K_{mi} the key chain of mention m_i, the precision and recall of the mention m_i are calculated as follows:

$$Precision(m_i) = \frac{|R_{mi} \bigcap K_{mi}|}{|R_{mi}|} \tag{2.4}$$

$$Recall(m_i) = \frac{|R_{mi} \bigcap K_{mi}|}{|K_{mi}|} \tag{2.5}$$

The overall precision and recall are computed by averaging them over all mentions. Figure 2.1 illustrates how B^3 scores are calculated given the key= $\{m_{1-5}\}$, $\{m_{6-7}\}$, $\{m_{8-12}\}$ and the system response= $\{m_{1-5}\}$, $\{m_{6-12}\}$.

2.4.3 CEAF

CEAF (***C**onstrained **E**ntity **A**ligned **F**-measure*) is developed by Luo [32] stands for . Luo criticizes the B3 algorithm for using entities more than one time, because

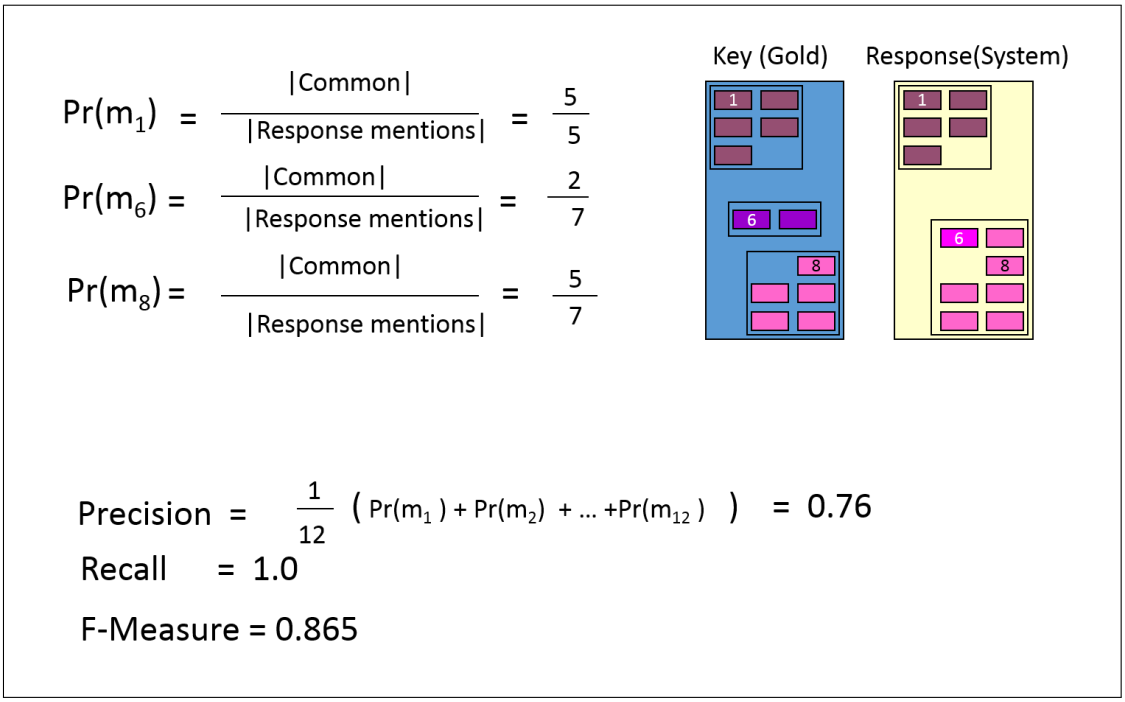

Figure 2.1 – Example on calculating B^3 metric scores

B3 computes precision and recall of mentions by comparing entities containing that mention. Thus, he proposed a new method based on entities instead of mentions. Here R_i is a system coreference chain and K_i is a key chain.

$$Precision = \frac{\phi(g^*)}{\sum_i \phi(R_i, R_i)} \tag{2.6}$$

$$Precision = \frac{\phi(g^*)}{\sum_i \phi(K_i, K_i)} \tag{2.7}$$

Where $\phi(g^*)$ is calculated as follow:

$$\phi(g^*) = max \begin{cases} \phi_3(K_i, R_j) = \left|K_i \bigcap R_j\right| \\ \\ \phi_4(K_i, R_j) = \frac{2\left|K_i \bigcap R_j\right|}{|K_i| + \left|R_j\right|} \end{cases} \tag{2.8}$$

Let suppose that we have:

Key = {a,b,c}

Response = {a,b,d}

$\phi_3(K_1, R_1) = 2(K_1 : \{a,b,c\}; R_1 : \{a,b,d\})$

$\phi_3(K_1, k_1) = 3$

$\phi_3(R_1, R_1) = 3$

The CEAF precision, recall and F-score for the example are calculated as:

$Precision = \frac{2}{3} = 0.667$

$Recall = \frac{2}{3} = 0.667$

$F1 = \frac{2 \cdot 0.667 \cdot 0.667}{0.67 + 0.667} = 0.667$

2.4.4 BLANC

BLANC [64] (for ***BiLa**teral **A**ssessment of **N**oun-phrase **C**oreference*) is the most recent introduced measure into the literature. This measure implements the Rand index [60] which has been originally developed to evaluate clustering methods. BLANC was mainly developed to deal with imbalance between singletons and coreferent mentions by considering coreference and non-coreference links. Figure 2.2 illustrates a gold (key) reference and the system response. First BLANC generate all possible mention pair combinations, calculated as follows:

$L = N * (N-1)/2$, where N is the number of mentions in the document.

Then it goes through each mention pair and classifies it in one of table 2.II four categories: **rc** : the number of **r**ight **c**oreference links (where both key and response say that the mention pair is coreferent); **wc**: the number of **w**rong **c**oreference links; **rn**: the number of **r**ight **n**on-coreference links; **wn**: the number of **w**rong **n**on-coreference links. In our example, rc = {m5-m12, m7-m9}, wc={m4-m6, m7-m14, m9-m14}, wn={m5-m14,m12-m14} and rn={The 84 right non-coreference mention pairs}.

Then, these values are filled in formulas of Table 2.III in order to calculate the final

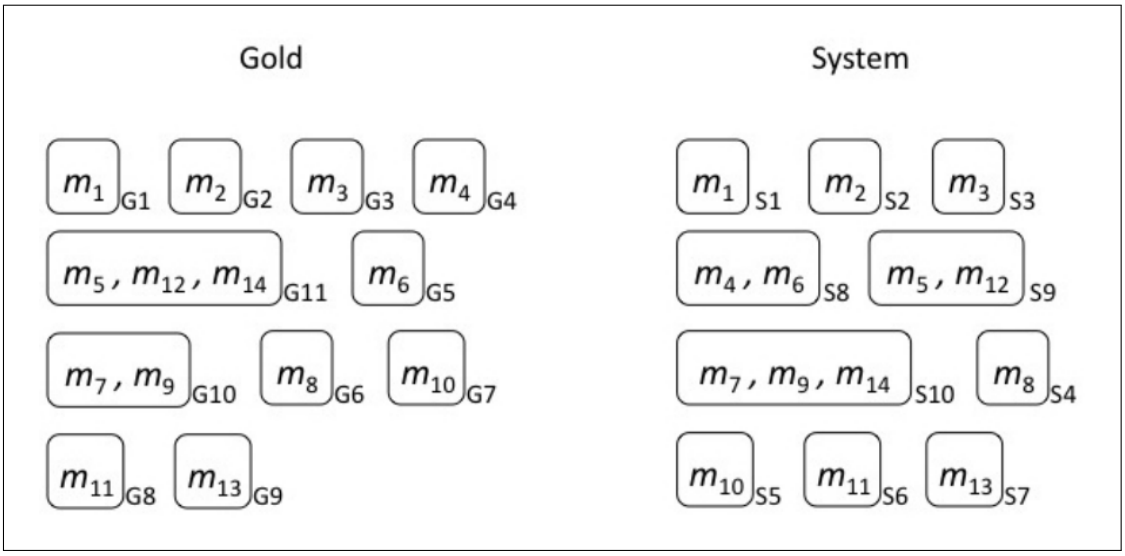

Figure 2.2 – Example of key (gold) and response (System) coreference chains

BLANC score. BLANC differs from other metrics by taking in consideration singleton clusters in the document, and crediting the system when it correctly identifies singleton instances. Consequently coreference links and non-coreference predictions contribute evenly in the final score.

2.4.5 CoNLL score and state-of-the-art Systems

This score is the average of MUC, B^3 , and CEAFϕ_4 F1. It was the official metric to determine the winning system in the CoNLL shared tasks of 2011 [54] and 2012 [55]. The CoNLL shared tasks of 2011 consist of identifying coreferring mentions in the English language portion of the OntoNotes data. Table 2.IV reports results of the top five systems that participated in the close track [3].

The task of 2012 extends the previous task by including data for Chinese and Arabic, in addition to English. After 2012, all works on coreference resolution adopt the official CoNLL train/test split in order to train and compare results. The last few years have seen a boost of work devoted to the development of machine learning based coreference

3. Full resluts can be found at `http://conll.cemantix.org/2011/`

		Response		Sum
		Coreference	Non-coreference	
KEY	Coreference	rc (2)	wn (2)	rc+wn (4)
	Non-coreference	wc (3)	rn (84)	wc+rn (87)
Sum		rc+wc (5)	wn+rn (86)	L (91)

Table 2.II – The BLANC confusion matrix, the values of example of Figure 2.2 are placed between parenthesizes.

Score	Coreference	Non-coreference	
P	$P_c = \frac{rc}{rc+wc}$	$P_n = \frac{rn}{rn+wn}$	$BLANC-P = \frac{P_c+P_n}{2}$
R	$R_c = \frac{rc}{rc+wn}$	$R_n = \frac{rn}{rn+wc}$	$BLANC-R = \frac{R_c+R_n}{2}$
F	$F_c = \frac{2P_cR_c}{P_c+R_c}$	$F_n = \frac{2P_nR_n}{P_n+R_n}$	$BLANC = \frac{F_c+F_n}{2}$

Table 2.III – Formula to calculate BLANC: precision recall and F1 score

System	MUC F^1	B^3 F^2	CEAFϕ_4 F^3	BLANC F	CoNLL $\frac{F^1+F^2+F^3}{3}$
lee	**59.57**	68.31	**45.48**	73.02	**57.79**
sapena	59.55	67.09	41.32	71.10	55.99
chang	57.15	**68.79**	41.94	**73.71**	55.96
nugues	58.61	65.46	39.52	71.11	54.53
santos	56.56	65.66	37.91	69.46	53.41

Table 2.IV – Performance of the top five systems in the CoNLL-2011 shared task

resolution systems. Table 2.V lists the performance of state-of-the-art systems (mid-2016) as reported in their respective paper .

System	MUC			B^3			CEAFϕ_4			CoNLL
	P	R	F1	P	R	F1	P	R	F1	F1
B&K (2014)	74.30	67.46	70.72	62.71	54.96	58.58	59.40	52.27	55.61	61.63
M&S (2015)	76.72	68.13	72.17	66.12	54.22	59.58	59.47	52.33	55.67	62.47
C&M (2015)	76.12	69.38	72.59	65.64	56.01	60.44	59.44	58.92	56.02	63.02
Wiseman et al. (2015)	76.23	69.31	72.60	66.07	55.83	60.52	59.41	54.88	57.05	63.39
Wiseman et al. (2016)	**77.49**	**69.75**	**73.42**	**66.83**	**56.95**	**61.50**	**62.14**	**53.85**	**57.70**	**64.21**

Table 2.V – Performance of current state-of-the-art systems on CoNLL 2012 English test set, including in order: [5]; [35]; [11]; [73] ; [74]

2.4.6 Wikipedia and Freebase

2.4.6.1 Wikipedia

Wikipedia is a very large domain-independent encyclopedic repository. The English version, as of 13 April 2013, contains 3,538,366 articles thus providing a large coverage knowledge resource.

Figure 2.3 – Excerpt from the Wikipedia article *Barack Obama*

An entry in Wikipedia provides information about the concept it mainly describes. A Wikipedia page has a number of useful reference features, such as: internal link or hyperlinks: link a surface form (*Label* in figure 2.3) into other article (*Wiki Article* in

figure 2.3) in Wikipedia); redirects: consist of misspelling and names variations of the article title; infobox: are structured information about the concept being described in the page; and categories: a semantic network classification.

2.4.6.2 Freebase

The aim of Freebase was to structure the human knowledge into a scalable tuple database, thus by collecting structured data from the web, where Wikipedia structured data (infobox) forms the skeleton of Freebase. As a result, each Wikipedia article has an equivalent page in Freebase, which contains well structured attributes related to the topic being described. Figure 2.4 shows some structured data from the Freebase page of *Barack Obama*.

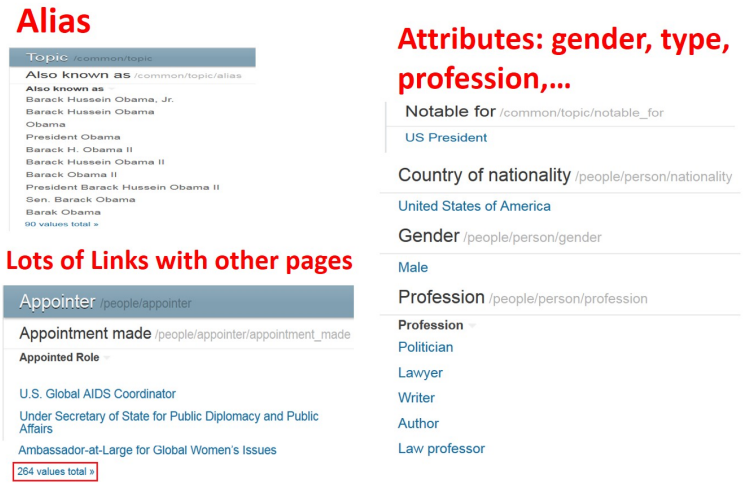

Figure 2.4 – Excerpt of the Freebase page of *Barack Obama*

CHAPTER 3

WIKICOREF: AN ENGLISH COREFERENCE-ANNOTATED CORPUS OF WIKIPEDIA ARTICLES

3.1 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 [22] to rule based systems [31], or the large-scale of annotated corpora comprising different text genres and languages.

Wikipedia [1] is a very large multilingual, domain-independent encyclopedic repository. The English version 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 [2] (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. Despite 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 ob-

1. https://www.wikipedia.org/
2. https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

jective to balance topics 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 [57], yet with some minor differences.

Contrary to similar endeavours discussed in Chapter 2, 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 this chapter is organized as follows. We describe the annotation process in Section 3.2. In Section 3.3, we present our annotation scheme along with a detailed description of attributes assigned to each mention. We present in Section 3.4 the main statistics of our corpus. Annotation reliability is measured in Section 3.5, before ending the chapter with conclusions and future works.

3.2 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.

3.2.1 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 (Figure 3.1) 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].

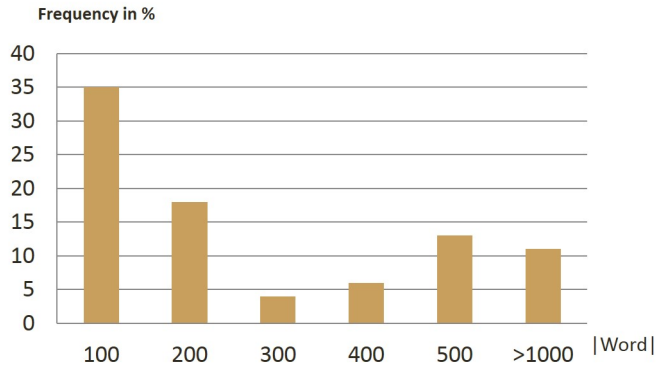

Figure 3.1 – Distribution of Wikipedia article depending on word count

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*.

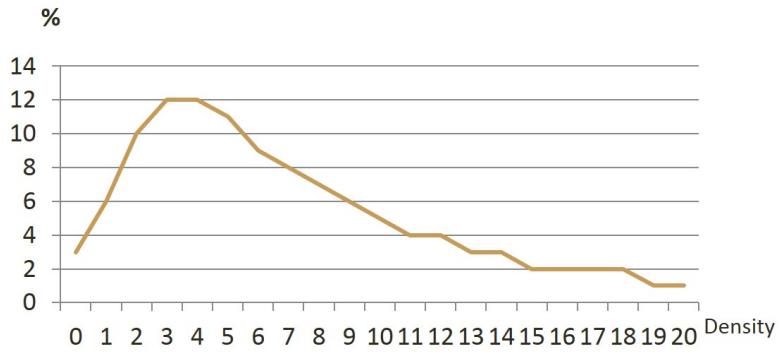

Figure 3.2 – Distribution of Wikipedia article depending on link density

In order to ensure that our corpus covers many topics of interest, we used the gazetteer generated by [61]. It contains a collection of 16 (high precision low recall) lists of Wikipedia article titles that cover diverse topics. It includes: Locations, Corporations, Occupations, Country, Man Made Object, Jobs, Organizations, Art Work, People, Competitions, Battles, Events, Place, Songs, Films. We selected our articles from all those lists, proportional to lists size.

3.2.2 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` [39]. 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.

3.2.3 Markables Extraction

We used the `Stanford CoreNLP` toolkit [34], 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 [31, 58], 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 advantage 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.

(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) Obama won **[**Best Spoken Word Album Grammy Awards**]** for abridged audio-book versions of **[**Dreams from My Father**]** ...

Figure 3.3 – Example of mentions detected by our method.

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 [6] aliases. When available in the Freebase 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`.

(a) He signed into law **[**the Car Allowance Rebate System**]**$_{\mathbf{X}}$, known colloquially as [“Cash for Clunkers”**]**$_{\mathbf{X}}$, that temporarily boosted the economy.

(b) ... the national holiday from Dominion Day to **[**Canada Day**]**$_{\mathbf{X}}$ in 1982 .... the 1867 Constitution Act officially proclaimed Canadian Confederation on **[**July 1 , 1867**]**$_{\mathbf{X}}$

Figure 3.4 – Example of mentions linked by our method.

All mentions that we detect this way allow us to extend `Dcoref` candidate list by mentions missed by the system (Fig.3.3). Also, all mentions that refer to the same

concept were linked into one coreference chain as in Fig.3.4. 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 larger Wikipedia\Freebase mention exists, as in Fig.3.5, or a Wikipedia\Freebase mention shares some content words with a mention detected by `Dcoref` (Fig.3.6).

(a) In December 2008, Time magazine named Obama as its **[**Person of <the Year>$_{\text{Dcoref}}$**]**$_{\text{Wiki/FB}}$ for his historic candidacy and election, which it described as "the steady march of seemingly impossible accomplishments".

(b) In a February 2009 poll conducted in Western Europe and the U.S. by Harris Interactive for **[**<France>$_{\text{Dcoref}}$ 24**]**$_{\text{Wiki/FB}}$

(c) He ended plans for a return of human spaceflight to the moon and development of **[**the Ares <I>$_{\text{Dcoref}}$ rocket**]**$_{\text{Wiki/FB}}$, **[**Ares <V>$_{\text{Dcoref}}$ rocket**]**$_{\text{Wiki/FB}}$

(d) His concession speech after the New Hampshire primary was set to music by independent artists as the music video **[**"Yes <We>$_{\text{Dcoref}}$ Can"**]**$_{\text{Wiki/FB}}$

Figure 3.5 – 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 [58]. We do so to avoid as much as possible noisy mentions in the pre-annotation phase.

(a) 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]$_{\text{Wiki/FB}}$, neither of which was signed into law>$_{\text{Dcoref}}$.

(b) Obama also sponsored a Senate amendment to **[**<the State Children's>$_{\text{Dcoref}}$ Health Insurance Program]$_{\text{Wiki/FB}}$

(c) In December 2006, President Bush signed into law the **[**Democratic Republic of the <Congo]$_{\text{Wiki/FB}}$ Relief>$_{\text{Dcoref}}$, Security, and Democracy Promotion Act

(d) Obama issued executive orders and presidential memoranda directing **[**<the U.S.>$_{\text{Dcoref}}$ military]$_{\text{Wiki/FB}}$ to develop plans to withdraw troops from Iraq.

Figure 3.6 – Examples of contradictions between Dcoref mentions (marked by angular brackets) and our method (marked by squared brackets)

Overall, we corrected roughly 15% of the 18212 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.

3.2.4 Annotation Tool and Format

Manual annotation is performed using `MMAX2` [41], which supports stand-off format. The toolkit allows multi-coding layers annotation at the same time and the graphical interface (Figure 3.7) introduces a multiple 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 (shown in Figure 3.8).

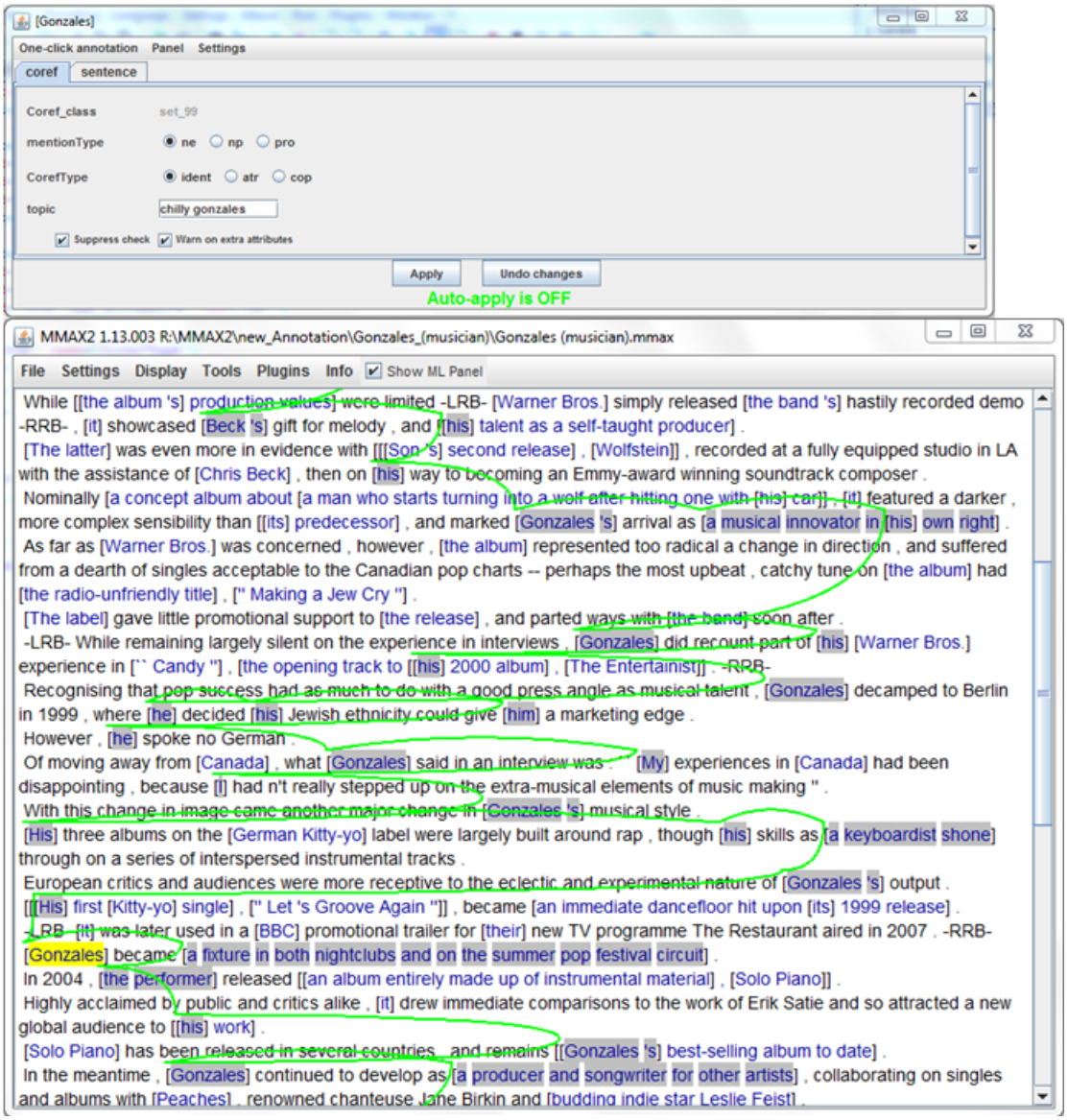

Figure 3.7 – Annotation of WikiCoref in MMAX2 tool

3.3 Annotation Scheme

In general, the annotation scheme in WikiCoref mainly follows the OntoNotes scheme [57]. In particular, only noun phrases are eligible to be mentions and only non-singleton coreference sets (coreference chain containing more than one mention) are kept in the

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml.org/NameSpaces/coref">
<markable id="markable_317" span="word_1..word_2"
coref_class="set_63" topic="
http://rdf.freebase.com/ns/m.0m34" coreftype="ident"
mentiontype="ne"  mmax_level="coref" />

<markable id="markable_323" span="word_13..word_13"
coref_class="set_63" topic="
http://rdf.freebase.com/ns/m.0m34" coreftype="ident"
mentiontype="pro"  mmax_level="coref" />

<markable id="markable_330" span="word_17..word_17"
coref_class="set_4" topic="
http://rdf.freebase.com/ns/m.05qtj" coreftype="ident"
mentiontype="ne"  mmax_level="coref" />

<markable id="markable_324" span="word_24..word_24"
coref_class="set_63" topic="
http://rdf.freebase.com/ns/m.0m34" coreftype="ident"
mentiontype="pro"  mmax_level="coref" />
```

Figure 3.8 – The XML format of the MMAX2 tool

version distributed. Each annotated mention is tagged by a set of attributes: mention type (Section 3.3.1), coreference type (Section 3.3.2) and the equivalent Freebase topic when available (Section 3.3.3). In Section 3.3.4, we introduce a few modifications we made to the OntoNotes guidelines in order to reduce ambiguity, consequently optimize our inter-annotator agreement.

3.3.1 Mention Type

3.3.1.1 Named entity (NE)

NEs can be proper names, noun phrases 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. The same rules is applied on dates, we illustrate this in the following example:

In a report issued January 5, 1995, the program manager said that there would be no new funds this year.

There is no relation to be marked between *1995* and *this year*, because the first mention is part of the larger NE *January 5, 1995*. 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. Similarly the 's is included in the NE span, as in *Groupe AG 's* chairman.

3.3.1.2 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*.

3.3.1.3 Pronominal (PRO)

Mentions tagged PRO may be one of the following subtypes:

Personal Pronouns: I, you, he, she, they, it excluding pleonastic it, me, him, us,

them, her and we.

Possessive Pronouns: my, your, his, her, its, mine, hers, our, your, their, ours, yours and theirs. In case that a reflexive pronoun is directly preceded by its antecedent, mentions are annotated as in the following example: *heading for mainland China or visiting [Macau [itself]$_X$]$_X$*.

Reflexive Pronouns: myself, yourself, himself, herself, themselves, itself, ourselves, yourselves and themselves.

Demonstrative Pronouns: this, that, these and those.

3.3.2 Coreference Type

MUC and ACE schemes treat identical (anaphor) and attributive (apositive or copular structure, see figure 3.9) mentions as coreferential, contrary to the OntoNotes scheme which differentiates between these two because they play different roles.

(a) **[**Jefferson Davis]$_{\text{ATR}}$, **[**President of the Confederate States of America]$_{\text{ATR}}$

(b) **[**The Prime Minister's Office]$_{\text{ATR}}$ ([PMO] $_{\text{ATR}}$) .

(c) a market value of **[**about 105 billion Belgian francs]$_{\text{ATR}}$ (**[**$ 2.7 billion] $_{\text{ATR}}$)

(d) **[**The Conservative lawyer] $_{\text{ATR}}$ **[**John P. Chipman] $_{\text{ATR}}$

(e) Borden is **[**the chancellor of Queen's University] $_{\text{COP}}$

Figure 3.9 – Example of Attributive and Copular mentions

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 -a- of Fig. 3.9), parenthetical (example -b- and -c-) or role appositive (example -d-) relation; and lastly Copular (COP) attributive mentions in copular structures (example -e-). We

added the latest because it offers useful information for coreference systems. For our annotation task, metonymy and acronym are marked as coreferential, as in Figure 3.10.

Metonymy Britain 's ..... the government

Metonymy the White House ..... the administration

Acronym The U.S ..... the country

Figure 3.10 – Example of Metonymy and Acronym mentions

3.3.3 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 (the `topic` attribute in figure 3.8) 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.

3.3.4 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.

3.3.4.1 Maximal Extent

In our annotation, we identify the maximal extent of the mention, thus including all modifiers of the mention: 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] $_X$, who slap a security guard] $_X$*
- *[[a colorful array] $_X$ of magazines] $_X$*

3.3.4.2 Verbs

Our annotation scheme does not support verbs or NP referring to them inclusively, as in the following example: *Sales of passenger cars [grew]$_V$ 22%. [The strong growth]$_{NP}$ followed year-to-year increases.*

3.4 Corpus Description

Corpus	**Size**	**#Doc**	**#Doc/Size**
ACE-2007 (English)	300k	599	500
[67]	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 3.I – Main characteristics of WikiCoref compared to existing coreference-annotated corpora

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

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

Table 3.II – Frequency of mention and coreference types in WikiCoref

The distribution of coreference and mentions types is presented in Table 3.II. 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 [67] the authors observe for instance that only 22.1% of mentions are named entities in their corpus of scientific articles.

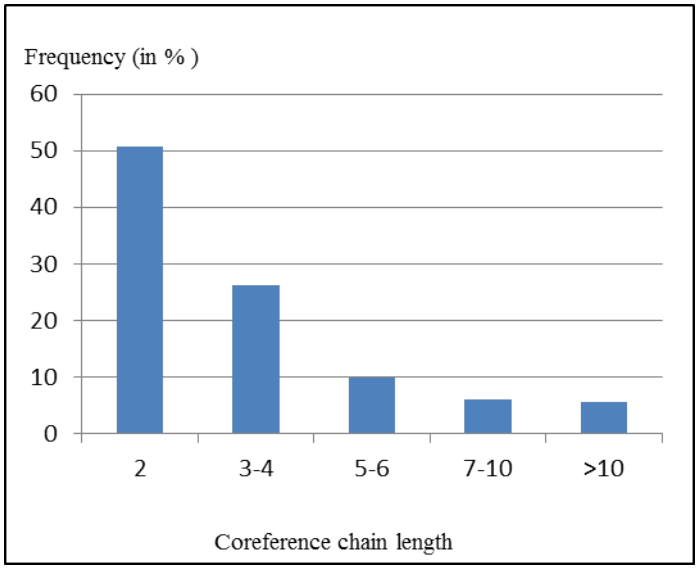

Figure 3.11 – Distribution of the coreference chains length

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 3.11. Also, WikiCoref contains 646 attributive mentions distributed over 330 attributive chains.

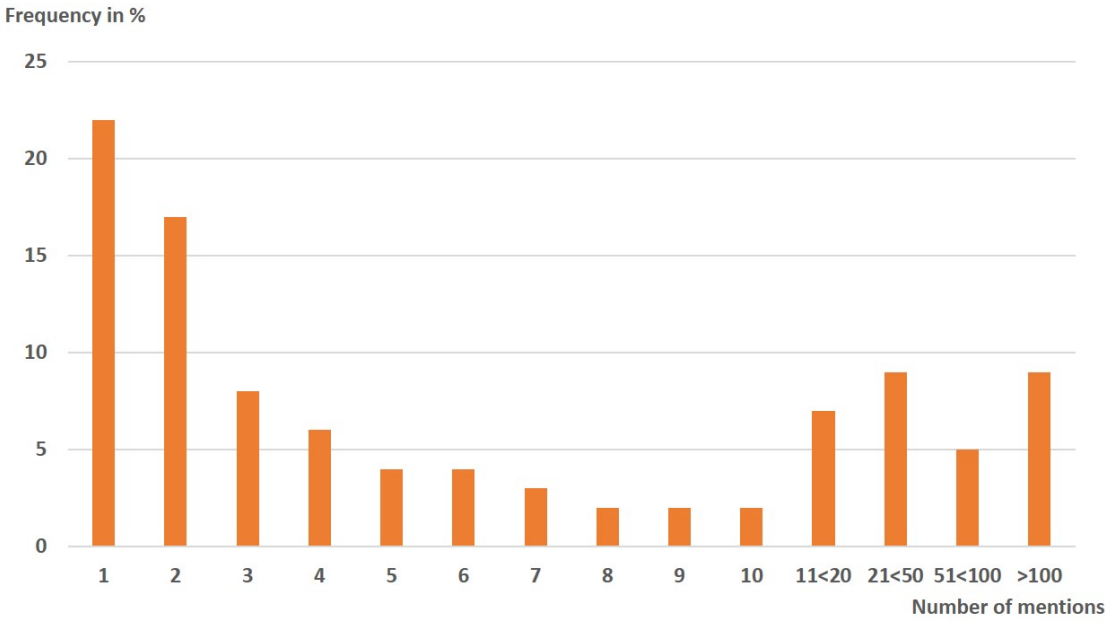

Figure 3.12 – Distribution of distances between two successive mentions in the same coreference chain

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. Figure 3.12 shows the number of mentions separating two successive mentions in the same coreference chain. Both distributions illustrated in Figures 3.11 and 3.12 apparently follow a curve of Zipfian type.

3.5 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 me. 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 [57] 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 [8]. The kappa coefficient calculate the agreement between annotators making category judgements, its calculated as follow:

$$K = \frac{P(A)-P(E)}{1-P(E)} \tag{3.1}$$

where P(A) is of times that annotators agree, and P(E) is the number of times that we expect that the annotators agree by chance. We reported a kappa of 0.78, which is very close to the well accepted threshold of 80%, but it falls in the range of other endeavours and it roughly indicates that both subjects often agreed.

We also measured a MUC F1 score [72] 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 [67] 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.

3.6 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 equiv-

alent 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.

CHAPTER 4

WIKIPEDIA MAIN CONCEPT DETECTOR

4.1 Introduction

Coreference Resolution (CR) is the task of identifying all mentions of entities in a document and grouping them into equivalence classes. CR is a prerequisite for many NLP tasks. For example, in Open Information Extraction (OIE) [79], one acquires subject-predicate-object relations, many of which (e.g., <the foundation stone, was laid by, the Queen 's daughter>) are useless because the subject or the object contains material coreferring to other mentions in the text being mined.

Most CR systems, including state-of-the-art ones [11, 20, 35] are essentially adapted to news-like texts. This is basically imputable to the availability of large datasets where this text genre is dominant. This includes resources developed within the Message Understanding Conferences (e.g., [25]) or the Automatic Content Extraction (ACE) program (e.g., [18]), as well as resources developed within the collaborative annotation project OntoNotes [57].

It is now widely accepted that coreference resolution systems trained on newswire data perform poorly when tested on other text genres [24, 67], including Wikipedia texts, as we shall see in our experiments.

Wikipedia is a large, multilingual, highly structured, multi-domain encyclopedia, providing an increasingly large wealth of knowledge. It is known to contain well-formed, grammatical and meaningful sentences, compared to say, ordinary internet documents. It is therefore a resource of choice in many NLP systems, see [36] for a review of some pioneering works.

Incorporating external knowledge into a CR system has been well studied for a number of years. In particular, a variety of approaches [22, 43, 53] have been shown to benefit from using external resources such as Wikipedia, WordNet [38], or YAGO [71]. [62] and [23] both investigate the integration of named-entity linking into machine learning

and rule-based coreference resolution system respectively. They both use GLOW [63] a *wikification* system which associates detected mentions with their equivalent entity in Wikipedia. In addition, they assign to each mention a set of highly accurate knowledge attributes extracted from Wikipedia and Freebase [6], such as the Wikipedia categories, gender, nationality, aliases, and NER type (ORG, PER, LOC, FAC, MISC).

One issue with all the aforementioned studies is that named entity linking is a challenging task [37], where inaccuracies often cause cascading errors in the pipeline [80]. Consequently, most authors concentrate on high-precision linking at the cost of low recall.

While Wikipedia is ubiquitous in the NLP community, we are not aware of much work conducted to adapt CR to this text genre. Two notable exceptions are [46] and [42], two studies dedicated to extract tuples from Wikipedia articles. Both studies demonstrate that the design of a dedicated rule-based CR system leads to improved extraction accuracy. The focus of those studies being information extraction, the authors did not spend much efforts in designing a fully-fledged CR designed for Wikipedia, neither did they evaluate it on a coreference resolution task.

Our main contribution in this work is to revisit the task initially discussed in [42] which consists in identifying in a Wikipedia article all the mentions of the concept being described by this article. We refer to this concept as the "main concept" (MC) henceforth. For instance, within the article `Chilly_Gonzales`, the task is to find all proper (e.g. *Gonzales*, *Beck*), nominal (e.g. *the performer*) and pronominal (e.g. *he*) mentions that refer to the MC "Chilly Gonzales".

For us, revisiting this task means that we propose a testbed for evaluating systems designed for it, and we compare a number of state-of-the-art systems on this testbed. More specifically, we frame this task as a binary classification problem, where one has to decide whether a detected mention refers to the MC. Our classifier exploits carefully designed features extracted from Wikipedia markup and characteristics, as well as from Freebase; many of which we borrowed from the related literature.

We show that our approach outperforms state-of-the-art generic coreference resolution engines on this task. We further demonstrate that the integration of our classifier

into the state-of-the-art rule-based coreference system of [31] improves the detection of coreference chains in Wikipedia articles.

The paper is organized as follows. We describe in Section 4.2 the baselines we built on top of two state-of-the-art coreference resolution systems, and present our approach in Section 4.3. We evaluate current state of the art system on WikiCoref in Section 4.4. We explain experiments we conducted on WikiCoref in section 4.5, and conclude in Section 4.6.

4.2 Baselines

Since there is no system readily available for our task, we devised four baselines on top of two available coreference resolution systems. Figure 4.1 illustrate the output of a CR system applied on the Wikipedia article *Barack Obama*. Our goal here is to isolate the coreference chain that represents the main concept (*Barack Obama* in this example).

c1 ∈ {Obama; his; he; I; He; Obama; Obama Sr.; He; President Obama; his}
c2 ∈{ the United States; the U.S.; United States }
c3 ∈{ Barack Obama; Obama , Sr.; he; His; Senator Obama }
c4 ∈{ John McCain; His; McCain; he}
c5 ∈{ Barack; he; me; Barack Obama}
c6 ∈{ Hillary Rodham Clinton; Hillary Clinton; her }
c7 ∈{ Barack Hussein Obama II; his}

Figure 4.1 – Output of a CR system applied on the Wikipedia article *Barack Obama*

We experimented with several heuristics, yielding the following baselines.

B1 picks the longest coreference chain identified and considers that its mentions are those that co-refer to the main concept. The baseline will select the chain **c1** as representative of the entity *Barack Obama* . The underlying assumption is that the most mentioned concept in a Wikipedia article is the main concept itself.

B2 picks the longest coreference chain identified if it contains a mention that exactly matches the MC title, otherwise it checks in decreasing order (longest to shortest) for a chain containing the title. This baseline will reject **c1** because it doesn't contain the exact title, so it will pick up **c3** as main concept reference. We expect this baseline to be more precise than the previous one overall.

As can be observed in figure 4.1, mentions of the MC often are spread over several coreference chains. Therefore we devised two more baselines that aggregate chains, with an expected increase in recall.

B3 conservatively aggregates chains containing a mention that exactly matches the MC title. The baseline will concatenate **c3** and **c5** to form the chain referring to *Barack Obama*.

B4 more loosely aggregates all chains that contain at least one mention whose span is a substring of the title [1]. For instance, given the main concept *Barack Obama*, we concatenate all chains containing either *Obama* or *Barack* in their mentions. In results, the output of this baseline will be **c1** + **c3** + **c5**. Obviously, this baseline should show a higher recall than the previous ones, but risks aggregating mentions that are not related to the MC. For instance, it will aggregate the coreference chain referring to *University of Sydney* concept with a chain containing the mention *Sydney*.

We observed that, for pronominal mentions, those baselines were not performing very well in terms of recall. With the aim of increasing recall, we added to the chain all the occurrences of pronouns found to refer to the MC (at least once) by the baseline. This heuristic was first proposed by [46]. For instance, if the pronoun *he* is found in the chain identified by the baseline, all pronouns *he* in the article are considered to be mentions of the MC *Barack Obama*. For example, the new baseline **B4** will contain along with mentions in **c1, c3 and c5**, the pronouns {*His; he*} from **c4** and {*his*} from **c7**. Obviously, there are cases where those pronouns do not co-refer to the MC, but this step significantly improves the performance on pronouns.

1. Grammatical words are not considered for matching.

4.3 Approach

Our approach is composed of a preprocessor which computes a representation of each mention in an article as well as its main concept; and a feature extractor which compares both representations for inducing a set of features.

4.3.1 Preprocessing

We extract mentions using the same mention detection algorithm embedded in [31] and [11]. This algorithm described in [58] extracts all named-entities, noun phrases and pronouns, and then removes spurious mentions.

We leverage the hyperlink structure of the article in order to enrich the list of mentions with shallow semantic attributes. For each link found within the article under consideration, we look through the list of predicted mentions for all mentions that match the surface string of the link. We assign to those mentions the attributes (entity type, gender and number) extracted from the Freebase entry (if it exists) corresponding to the Wikipedia article the hyperlink points to. This module behaves as a substitute to the named-entity linking pipelines used in other works, such as [23, 62]. We expect it to be of high quality because it exploits human-made links.

We use the `WikipediaMiner` [39] API for easily accessing any piece of structure (clean text, labels, internal links, redirects, etc) in Wikipedia, and Jena [2] to index and query Freebase.

In the end, we represent a mention by three strings, as well as its coarse attributes (entity type, gender and number). Figure 4.2 shows the representation collected for the mention *San Fernando Valley region of the city of Los Angeles* found in the `Los_Angeles_Pierce_College` article.

We represent the main concept of a Wikipedia article by its **title**, its **inferred type** (a common noun inferred from the first sentence of the article). Those attributes were used in [46] to heuristically link a mention to the main concept of an article. We further extend this representation by the MC **name variants** extracted from the markup

2. http://jena.apache.org

string span

▷ *San Fernando Valley region of the city of Los Angeles*

head word span

▷ *region*

span up to the head noun

▷ *San Fernando Valley region*

coarse attribute

▷ ∅*, neutral, singular*

Figure 4.2 – Representation of a mention.

of Wikipedia (redirects, text anchored in links) as well as aliases from Freebase; the MC **entity types** we extracted from the Freebase `notable types` attribute, and its **coarse attributes** extracted from Freebase, such as its NER type, its gender and number. If the concept category is a person (PER), we import the `profession` attribute. Figure 4.3 illustrates the information we collect for the Wikipedia concept `Los_Angeles_Pierce_College`.

4.3.2 Feature Extraction

We experimented with a few hundred features for characterizing each mention, focusing on the most promising ones that we found simple enough to compute. In part, our features are inspired by coreference systems that use Wikipedia and Freebase as feature sources. These features, along with others related to the characteristics of Wikipedia texts, allow us to recognize mentions of the MC more accurately than current CR systems. We make a distinction between features computed for pronominal mentions and features computed from the other mentions.

4.3.2.1 Non-pronominal Mentions

For each mention, we compute seven families of features we describe below.

title	(W)
▷ *Los Angeles Pierce College*	
inferred type	(W)
Los Angeles Pierce College, also known as Pierce College and just Pierce, is a community college that serves . . .	
▷ *college*	
name variants	(W,F)
▷ *Pierce Junior College*, *LAPC*	
entity type	(F)
▷ College/University	
coarse attributes	(F)
▷ ORG, neutral, singular	

Figure 4.3 – Representation of a Wikipedia concept. The source from which the information is extracted is indicated in parentheses: (W)ikipedia, (F)reebase.

base Number of occurrences of the mention span and the mention head found in the list of candidate mentions. We also add a normalized version of those counts (frequency / total number of mentions in the list).

title, inferred type, name variants, entity type Most often, a concept is referred to by its name, one of its variants, or its type which are encoded in the four first fields of our MC representation. We define four families of comparison features, each corresponding to one of the first four fields of a MC representation (see Figure 4.3). For instance, for the title family, we compare the title text span with each of the text spans of the mention representation (see Figure 4.2). A comparison between a field of the MC representation and a mention text span yields 10 boolean features. These features encode string similarities (exact match, partial match, one being the substring of another, sharing of a number of words, etc.). An eleventh feature is the semantic relatedness score of [76]. For **title**, we

therefore end up with 3 sets (*titleSpan_MentionSpan*, *titleSpan_MentionHead* and *titleSpan_MentionSpanUpToHead*) of 11 feature vectors (illustrated in Figure 4.I).

Feature	**MC String**	**Mention String**
Equal	Pierce Junior College	Pierce Junior College
Equal Ignore Case	Pierce Junior College	Pierce junior college
Included in	College	Pierce College
Included in Ignore Case	college	Pierce College
Domain	Clarence W. Pierce School of Agriculture	Pierce
Domain Ignore Case	Clarence W. Pierce School of Agriculture	school
MC starts with Mention	Los Angeles Pierce College	Los Angeles
MC ends with Mention	Los Angeles Pierce College	Pierce College
Mention starts with MC	college	the college farm
Mention ends with MC	College	Pierce College
WordNet Sim. = 0.625	college	school

Table 4.I – The eleven feature encoding string similarity (10 row) and semantic similarity (row number 11). Columns two and three contain possible values of strings representing the MC (title or alias...) and a mention (mention span or head...) respectively. The last row shows the WordNet similarity between MC and mention strings.

tag Part-of-speech tags of the first and last words of the mention, as well as the tag of the words immediately before and after the mention in the article. We convert this into 34×4 binary features (presence/absence of a specific combination of tags).

main Boolean features encoding whether the MC and the mention **coarse attributes** match. Table 4.II illustrates matching between attributes of the MC (*Los Angeles Pierce College*) and the mention (*Los Angeles*) reconized by our preprocessing method as a referent of "The city of Los Angeles". Also we use conjunctions of all pairs of features in this family.

Feature	MC	Mention	Value
entity type	ORG	LOC	False
gender	neutral	neutral	true
number	singular	singular	true

Table 4.II – The non-pronominal mention *main* features family

4.3.2.2 Pronominal Mentions

We characterize pronominal mentions by five families of features, which, with the exception of the first one, all capture information extracted from Wikipedia.

base The pronoun span itself, number, gender and person attributes, to which we add the number of occurrences of the pronoun, as well as its normalized count. The most frequently occurring pronoun in an article is likely to co-refer to the main concept, and we expect these features to capture this to some extent.

main MC coarse attributes, such as NER type, gender, number (see Figure 4.3). That is, we use only those three values as features without conjoining them with the mention attributes as in non-pronominal features.

tag Part-of-speech of the previous and following tokens, as well as the previous and the next POS bigrams (this is converted into 2380 binary features).

position Often, pronouns at the beginning of a new section or paragraph refer to the main concept. Therefore, we compute 4 (binary) features encoding the relative position (first, first tier, second tier, last tier, last) of a mention in the sentence, paragraph, section and article.

distance Within a sentence, we search before and after the mention for an entity that is compatible (according to Freebase information) with the pronominal mention of interest. If a match is found, one feature encodes the distance between the match and the mention; another feature encodes the number of other compatible pronouns in the same sentence. We expect that this family of features will help the model to capture the presence of local (within a sentence) co-references.

4.4 Dataset

As our approach is dedicated to Wikipedia articles, we used WikiCoref described in chapter 3. Since most coreference resolution systems for English are trained and tested on ACE [18] or OntoNotes [27] resources, it is interesting to measure how state-of-the art systems perform on the WikiCoref dataset. To this end, we ran a number of recent CR systems: the rule-based system of [31] we call it `Dcoref`; the Berkeley systems described in [19, 20]; the latent model of [35] we call it `Cort` in Table 4.III; and the system described in [11] we call it `Scoref` which achieved the best results to date on the CoNLL 2012 test set.

System	WikiCoref	OntoNotes
`Dcoref`	51.77	55.59
[19]	51.01	61.41
[20]	49.52	61.79
`Cort`	49.94	62.47
`Scoref`	46.39	63.61

Table 4.III – CoNLL F1 score of recent state-of-the-art systems on the WikiCoref dataset, and the 2012 OntoNotes test data for predicted mentions.

We evaluate the systems on the whole dataset, using the v8.01 of the CoNLL scorer [3] [56]. The results are reported in Table 4.III along with the performance of the systems on the CoNLL 2012 test data [55]. Expectedly, the performance of all systems dramatically decrease on WikiCoref, which calls for further research on adapting the coreference resolution technology to new text genres. What is more surprising is that the rule-based system of [31] works better than the machine-learning based systems on the WikiCoref dataset, note however that we didn't train those systems on WikiCoref. Also, the ranking of the statistical systems on this dataset differs from the one obtained on the OntoNotes test set.

3. `http://conll.github.io/reference-coreference-scorers`

We believe our results to be representative, even if WikiCoref is smaller than the widely used OntoNotes. Those results further confirm the conclusions in [24], which show that a CR system trained on news-paper significantly underperforms on data coming from users comments and blogs. Nevertheless, statistical systems can be trained or adapted to the WikiCoref dataset, a point we leave for future investigations.

We generated baselines for all the systems discussed in this section, results are in table 4.V.

4.5 Experiments

In this section, we first describe the data preparation we conducted (section 4.5.1), and provide details on the classifier we trained (section 4.5.2). Then, we report experiments we carried out on the task of identifying the mentions co-referent (positive class) to the main concept of an article (section 4.5.3). We compare our approach to the baselines described in section 4.2, and analyze the impact of the families of features described in section 4.3. We also investigate a simple extension of `Dcoref` which takes advantage of our classifier for improving coreference resolution (section 4.5.4).

4.5.1 Data Preparation

Each article in WikiCoref was part-of-speech tagged, syntactically parsed and the named-entities were identified. This was done thanks to the `Stanford CoreNLP` toolkit [34]. Since WikiCoref does not contain singleton mentions (in conformance to the OntoNotes guidelines), we consider the union of WikiCoref mentions and all mentions predicted by the method described in [58]. Overall, we added about 13 400 automatically extracted mentions (singletons) to the 7 000 coreferent mentions annotated in WikiCoref. In the end, our training set consists of 20 362 mentions: 1 334 pronominal ones (627 of them referring to the MC), and 19 028 non-pronominal ones (16% of them referring to the MC).

4.5.2 Classifier

We trained two Support Vector Machine classifiers [13], one for pronominal mentions and one for non-pronominal ones, making use of the LIBSVM library [10] and the features described in Section 4.3.2. For both models, we selected[4] the C-support vector classification and used a linear kernel. Since our dataset is unbalanced (at least for non-pronominal mentions), we penalized the negative class with a weight of 2.0. Configuration of the SVM used in this experiment are in Table 4.IV.

Parameter	**Value**
Cachesize	40
kernel Type	Linear
SVM Type	C-SVC
Coef0	0
Cost	1.0
Shrinking	False
Weight	2.0 1.0

Table 4.IV – Configuration of the SVM classifier for both pronominal and non pronominal models

During training, we do not use gold mention attributes, but we automatically enrich mentions with the information extracted from Wikipedia and Freebase, as described in Section 4.3.

4. We tried with less success other configurations on a held-out dataset.

System	Pronominal			Non Pronominal			All		
	P	R	F1	P	R	F1	P	R	F1
Dcoref									
B1	64.51	76.55	70.02	70.33	63.09	66.51	67.92	67.77	67.85
B2	76.45	50.23	60.63	83.52	49.57	62.21	80.90	49.80	61.65
B3	76.39	65.55	70.55	83.67	56.20	67.24	80.72	59.45	68.47
B4	71.74	83.41	77.13	74.39	75.59	74.98	73.30	78.31	75.77
D&K (2013)									
B1	64.81	92.82	76.32	76.51	55.95	64.63	70.53	68.77	69.64
B2	80.94	79.26	80.09	90.78	52.8	66.77	86.13	62.0	72.1
B3	78.64	81.65	80.12	90.26	59.94	72.04	84.98	67.49	75.23
B4	72.09	**93.93**	81.57	78.28	65.9	71.56	75.48	75.65	75.56
D&K (2014)									
B1	65.23	87.08	74.59	70.59	36.13	47.8	67.47	53.85	59.9
B2	83.66	53.11	64.97	87.57	26.36	40.52	85.5	35.66	50.33
B3	81.3	77.67	79.44	83.28	52.12	64.12	82.39	61.0	70.1
B4	72.13	93.30	81.36	73.72	67.77	70.62	73.04	76.65	74.8
Cort									
B1	69.65	87.87	77.71	64.05	38.94	48.43	66.99	55.96	60.98
B2	89.57	67.14	76.75	80.91	33.16	47.04	85.18	44.98	58.87
B3	81.89	74.32	77.92	79.46	55.95	65.66	80.45	62.34	70.25
B4	77.36	89.95	83.18	71.51	67.26	69.32	73.84	75.15	74.49
Scoref									
B1	76.59	78.30	77.44	54.66	39.37	45.77	64.11	52.91	57.97
B2	89.59	74.16	81.15	69.90	31.20	43.15	79.69	46.14	58.44
B3	83.91	77.35	80.49	73.17	55.44	63.08	77.39	63.06	69.49
B4	78.48	90.74	84.17	67.51	67.85	67.68	71.68	75.81	73.69
this work	**85.46**	92.82	**88.99**	**91.65**	**85.88**	**88.67**	**89.29**	**88.30**	**88.79**

Table 4.V – Performance of the baselines on the task of identifying all MC coreferent mentions.

4.5.3 Main Concept Resolution Performance

We focus on the task of identifying all the mentions referring to the main concept of an article. We measure the performance of the systems we devised by average precision, recall and F1 rates computed by a 10-fold cross-validation procedure.

The results of the baselines and our approach are reported in Table 4.V. Clearly, our approach outperforms all baselines for both pronominal and non-pronominal mentions, and across all metrics. On all mentions, our best classifier yields an absolute F1 increase of 13 points over the best baseline (B4 of `Dcoref`).

In order to understand the impact of each family of features we considered in this study, we trained various classifiers in a greedy fashion. We started with the simplest feature set (**base**) and gradually added one family of features at a time, keeping at each iteration the one leading to the highest increase in F1. The outcome of this process for the pronominal mentions is reported in Table 4.VI.

	P	R	F1
always positive	46.70	100.00	63.70
base	70.34	78.31	74.11
+main	74.15	90.11	81.35
+position	80.43	89.15	84.57
+tag	82.12	90.11	85.93
+distance	85.46	92.82	88.99

Table 4.VI – Performance of our approach on the pronominal mentions, as a function of the features.

A baseline that always considers that a pronominal mention is co-referent to the main concept results in an F1 measure of 63.7%. This naive baseline is outperformed by the simplest of our model (**base**) by a large margin (over 10 absolute points). We observe that recall significantly improves when those features are augmented with the MC coarse attributes (**+main**). In fact, this variant already outperforms all the `Dcoref`-based baselines in terms of F1 score. Each feature family added further improves the

performance overall, leading to better precision and recall than any of the baselines tested.

Inspection shows that most of the errors on pronominal mentions are introduced by the lack of information on noun phrase mentions surrounding the pronouns. In example (f) shown in Figure 3, the classifier associates the mention *it* with the MC instead of *the Johnston Atoll " Safeguard C " mission*.

Table 4.VII reports the results obtained for the non-pronominal mentions classifier. The simplest classifier is outperformed by most baselines in terms of F1. Still, this model is able to correctly match mentions in example (a) and (b) of Figure 4.4 simply because those mentions are frequent within their respective article. Of course, such a simple model is often wrong as in example (c), where all mentions *the United States* are associated to the MC, simply because this is a frequent mention.

	P	R	F1
base	60.89	62.24	61.56
+title	85.56	68.03	75.79
+inferred type	87.45	75.26	80.90
+name variants	86.49	81.12	83.72
+entity type	86.37	82.99	84.65
+tag	87.09	85.46	86.27
+main	91.65	85.88	88.67

Table 4.VII – Performance of our approach on the non-pronominal mentions, as a function of the features.

The **title** feature family drastically increases precision, and the resulting classifier (**+title**) outperforms all the baselines in terms of F1 score. Adding the **inferred type** feature family gives a further boost in recall (7 absolute points) with no loss in precision (gain of almost 2 points). For instance, the resulting classifier can link the mention *the team* to the MC *Houston Texans* (see example (d)) because it correctly identifies the term *team* as a type. The family **name variants** also gives a nice boost in recall, in

a slight expense of precision. This drop is due to some noisy redirects in Wikipedia, misleading our classifier. For instance, *Johnston and Sand Islands* is a redirect of the `Johnston_Atoll` article.

a	MC= *Anatole France* *France* is also widely believed to be the model for narrator Marcel's literary idol Bergotte in Marcel Proust's In Search of Lost Time.
b	MC= *Harry Potter and the Chamber of Secrets* Although Rowling found it difficult to finish *the book*, it won ....
c	MC= *Barack Obama* On August 31, 2010, Obama announced that *the United States** combat mission in Iraq was over.
d	MC= *Houston Texans* In 2002, *the team* wore a patch commemorating their inaugural season...
e	MC= *Houston Texans* The name Houston Oilers was unavailable to *the expansion team*...
f	MC= *Johnston Atoll* In 1993 , Congress appropriated no funds for the Johnston Atoll Safeguard C mission , bringing *it** to an end.
g	MC= *Houston Texans* The Houston Texans are a professional American football team based in *Houston** , Texas.

Figure 4.4 – Examples of mentions (underlined) associated with the MC. An asterisk indicates wrong decisions.

The **entity type** family further improves performance, mainly because it plays a role similar to the **inferred type** features extracted from Freebase. This indicates that the noun type induced directly from the first sentence of a Wikipedia article is pertinent and can complement the types extracted from Freebase when available or serve as proxy when they are missing. Finally, the **main** family significantly increases precision (over 4 absolute points) with no loss in recall. To illustrate a negative example, the resulting

classifier wrongly recognizes mentions referring to the town *Houston* as coreferent to the football team in example (g). We handpicked a number of classification errors and found that most of these are difficult coreference cases. For instance, our best classifier fails to recognize that the mention *the expansion team* refers to the main concept *Houston Texans* in example (e).

4.5.4 Coreference Resolution Performance

Identifying all the mentions of the MC in a Wikipedia article is certainly useful in a number of NLP tasks [42, 46]. Finding all coreference chains in a Wikipedia article is worth studying. In the following, we describe an experiment where we introduced in `Dcoref` a new high-precision sieve which uses our classifier [5]. Sieves in `Dcoref` are ranked in decreasing order of precision, and we ranked this new sieve first. The aim of this sieve is to construct the coreference chain equivalent to the main concept. It merges two chains whenever they both contain mentions to the MC according to our classifier. We further prevent other sieves from appending new mentions to the MC coreference chain.

System	MUC			B^3			CEAFϕ_4			CoNLL
	P	R	F1	P	R	F1	P	R	F1	F1
`Dcoref`	61.59	60.42	61.00	53.55	43.33	47.90	42.68	**50.86**	46.41	51.77
D&K (2013)	68.52	55.96	61.61	59.08	39.72	47.51	48.06	40.44	43.92	51.01
D&K (2014)	63.79	57.07	60.24	52.55	40.75	45.90	45.44	39.80	42.43	49.52
M&S (2015)	**70.39**	53.63	60.88	**60.81**	37.58	46.45	**47.88**	38.18	42.48	49.94
C&M (2015)	69.45	49.53	57.83	57.99	34.42	43.20	46.61	33.09	38.70	46.58
`Dcoref++`	66.06	**62.93**	**64.46**	57.73	**48.58**	**52.76**	46.76	49.54	**48.11**	**55.11**

Table 4.VIII – Performance of `Dcoref++` on WikiCoref compared to state of the art systems, including in order: [31]; [19] - Final; [20] - Joint; [35] - Ranking:Latent; [11] - Statistical mode with clustering.

We ran this modified system (called `Dcoref++`) on the WikiCoref dataset, where

5. We use predicted results from 10-fold cross-validation.

mentions were automatically predicted. The results of this system are reported in Table 4.VIII, measured in terms of MUC [72], B3 [2], CEAFϕ_4 [32] and the average F1 CoNLL score [16].

We observe an improvement for `Dcoref++` over the other systems, for all the metrics. In particular, `Dcoref++` increases by 4 absolute points the CoNLL F1 score. This shows that early decisions taken by our classifier benefit other sieves as well. It must be noted, however, that the overall gain in precision is larger than the one in recall.

4.6 Conclusion

We developed a simple yet powerful approach that accurately identifies all the mentions that co-refer to the concept being described in a Wikipedia article. We tackle the problem with two (pronominal and non-pronominal) models based on well designed features. The resulting system is compared to baselines built on top of state-of-the-art systems adapted to this task. Despite being relatively simple, our model reaches 89 % in F1 score, an absolute gain of 13 F1 points over the best baseline. We further show that incorporating our system into the Stanford deterministic rule-based system [31] leads to an improvement of 4% in F1 score on a fully fledged coreference task.

In order to allow other researchers to reproduce our results, and report on new ones, we share all the datasets we used in this study. We also provide a dump of all the mentions in English Wikipedia our classifier identified as referring to the main concept, along with information we extracted from Wikipedia and Freebase.

In this master thesis, we proposed an approach to solve the problem of identifying all the mentions of the main concept in its Wikipedia article. While the proposed approach showed improved results compared to the state-of-the-art, it opens the door to a range of new research directions for other NLP tasks, which could be studied in future work.

In this section we list a number of directions in which to extend the work presented here. We believe that the MC mentions are the key to transform Wikipedia into training data thus provides an alternative to the manual and expensive annotation required for several NLP tasks. One way to do so is by taking the non-pronominal mentions of a

source article (e.g. Obama, the president, Senator Obama for the article Barack Obama), and tracking those spans in a "target article", where the source appears as an internal hyperlink in the target article.

This approach is an extension to approaches found in the literature which use only human labelled links as training data for their respective tasks, such as Named Entity Recognition [49] and Entity Linking [70]. We believe that our method will add valuable annotations, consequently improving the performance of statistical NER/EL systems.

Another direction of future work is to integrate our classifier in OIE systems on Wikipedia which in turn will improve the quality of the extracted triples and save many of them which contain coreferential material. To the best of our knowledge, the impact of coreference resolution to OIE is an issue of IE that has never been studied. Finally, a natural extension of this work is to employ the MC mentions in order to identify all coreference relations in a Wikipedia article, a task we are currently investigating.

BIBLIOGRAPHY

[1] Hiyan Alshawi. Resolving quasi logical forms. *Computational Linguistics*, 16(3): 133–144, 1990.

[2] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566, 1998.

[3] Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303, 2008.

[4] Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, and Frank Rudzicz. Using knowledge-poor coreference resolution for text summarization. In *Proceedings of DUC*, volume 3, 2003.

[5] Anders Björkelund and Jonas Kuhn. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *ACL (1)*, pages 47–57, 2014.

[6] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 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, 2008.

[7] Jaime G Carbonell and Ralf D Brown. Anaphora resolution: a multi-strategy approach. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 96–101, 1988.

[8] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[9] José Castano, Jason Zhang, and James Pustejovsky. Anaphora resolution in biomedical literature. 2002.

[10] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[11] Kevin Clark and Christopher D. Manning. Entity-centric coreference resolution with model stacking. In *Association of Computational Linguistics (ACL)*, 2015.

[12] K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr, Christophe Roeder, Philip V. Ogren, Martha Palmer, and Lawrence Hunter. Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *Proceedings of BioTxtM 2010: 2nd workshop on building and evaluating resources for biomedical text mining*, pages 37–41, 2010.

[13] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[14] Aron Culotta, Michael Wick, Robert Hall, and Andrew McCallum. First-order probabilistic models for coreference resolution. 2006.

[15] Pascal Denis. *New learning models for robust reference resolution*. 2007.

[16] Pascal Denis and Jason Baldridge. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42(1):87–96, 2009.

[17] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1, 2004.

[18] George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *LREC*, volume 2, page 1, 2004.

[19] Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *EMNLP*, pages 1971–1982, 2013.

[20] Greg Durrett and Dan Klein. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490, 2014.

[21] Ralph Grishman. The nyu system for muc-6 or where's the syntax? In *Proceedings of the 6th conference on Message understanding*, pages 167–175, 1995.

[22] Aria Haghighi and Dan Klein. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1152–1161, 2009.

[23] Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke S. Zettlemoyer. Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves. In *EMNLP*, pages 289–299, 2013.

[24] Iris Hendrickx and Veronique Hoste. Coreference resolution on blogs and commented news. In *Anaphora Processing and Applications*, pages 43–53. Springer, 2009.

[25] Lynette Hirshman and Nancy Chinchor. MUC-7 coreference task definition. version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[26] Jerry R Hobbs. Resolving pronoun references. *Lingua*, 44(4):311–338, 1978.

[27] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics, 2006.

[28] Krippendorff Klaus. *Content analysis: An introduction to its methodology*. Sage Publications, 1980.

[29] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180, 2007.

[30] Shalom Lappin and Herbert J Leass. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561, 1994.

[31] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916, 2013.

[32] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics, 2005.

[33] Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 135, 2004.

[34] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

[35] Sebastian Martschat and Michael Strube. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418, 2015.

[36] Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. Mining meaning from wikipedia. *Int. J. Hum.-Comput. Stud.*, 67(9):716–754, September 2009.

[37] Rada Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. In *HLT-NAACL*, pages 196–203, 2007.

[38] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38 (11):39–41, 1995.

[39] David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[40] Dan I Moldovan, Sanda M Harabagiu, Roxana Girju, Paul Morarescu, V Finley Lacatusu, Adrian Novischi, Adriana Badulescu, and Orest Bolohan. Lcc tools for question answering. In *TREC*, 2002.

[41] Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214, 2006.

[42] Kotaro Nakayama. Wikipedia mining for triple extraction enhanced by co-reference resolution. In *The 7th International Semantic Web Conference*, page 103, 2008.

[43] Vincent Ng. Shallow Semantics for Coreference Resolution. In *IJcAI*, volume 2007, pages 1689–1694, 2007.

[44] Vincent Ng and Claire Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7, 2002.

[45] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111, 2002.

[46] Dat PT Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. Relation extraction from wikipedia using subtree mining. In *Proceedings of the National Conference on Artificial Intelligence*, page 1414, 2007.

[47] N. Nguyen, J. D. Kim, and J. Tsujii. Overview of bionlp 2011 protein coreference shared task. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 74–82, 2011.

[48] Nicolas Nicolov, Franco Salvetti, and Steliana Ivanova. Sentiment analysis: Does coreference matter. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 37, 2008.

[49] Joel Nothman, James R Curran, and Tara Murphy. Transforming wikipedia into named entity training data. In *Proceedings of the Australian Language Technology Workshop*, pages 124–132, 2008.

[50] Massimo Poesio. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79. Association for Computational Linguistics, 2004.

[51] Massimo Poesio, Barbara Di Eugenio, and Gerard Keohane. Discourse structure and anaphora: An empirical study. 2002.

[52] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, 1998.

[53] Simone Paolo Ponzetto and Michael Strube. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199, 2006.

[54] Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 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. Association for Computational Linguistics, 2011.

[55] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics, 2012.

[56] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, June 2014.

[57] Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. Unrestricted coreference: Identifying entities and events in OntoNotes. In *First IEEE International Conference on Semantic Computing*, pages 446–453, 2007.

[58] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics, 2010.

[59] Altaf Rahman and Vincent Ng. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 968–977, 2009.

[60] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[61] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named en-

tity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, 2009.

[62] Lev Ratinov and Dan Roth. Learning-based multi-sieve co-reference resolution with knowledge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1234–1244, 2012.

[63] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384, 2011.

[64] Marta Recasens and Eduard Hovy. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510, 2011.

[65] Elaine Rich and Susann LuperFoy. An architecture for anaphora resolution. In *Proceedings of the second conference on Applied natural language processing*, pages 18–24, 1988.

[66] Kepa Joseba Rodrıguez, Francesca Delogu, Yannick Versley, Egon W. Stemle, and Massimo Poesio. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proceedings of LREC*, pages 157–163. Citeseer, 2010.

[67] Ulrich Schäfer, Christian Spurk, and Jörg Steffen. A fully coreference-annotated corpus of scholarly papers from the acl anthology. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING-2012)*, pages 1059–1070, 2012.

[68] Isabel Segura-Bedmar, Mario Crespo, César de Pablo-Sánchez, and Paloma Martínez. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC bioinformatics*, 11(2):1, 2010.

[69] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27 (4):521–544, 2001.

[70] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424, 2006.

[71] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.

[72] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics, 1995.

[73] Sam Wiseman, Alexander M Rush, Stuart M Shieber, Jason Weston, Heather Pon-Barry, Stuart M Shieber, Nicholas Longenbaugh, Sam Wiseman, Stuart M Shieber, Elif Yamangil, et al. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 92–100, 2015.

[74] Sam Wiseman, Alexander M Rush, and Stuart M Shieber. Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*, 2016.

[75] Fei Wu and Daniel S. Weld. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, 2010.

[76] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[77] Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual*

Meeting on Association for Computational Linguistics-Volume 1, pages 176–183, 2003.

[78] Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, Ting Liu, and Sheng Li. An entity-mention model for coreference resolution with inductive logic programming. In *ACL*, pages 843–851, 2008.

[79] Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. Textrunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 25–26. Association for Computational Linguistics, 2007.

[80] Jianping Zheng, Luke Vilnis, Sameer Singh, Jinho D. Choi, and Andrew McCallum. Dynamic knowledge-base alignment for coreference resolution. In *Conference on Computational Natural Language Learning (CoNLL)*, 2013.