Master 2 Traitement automatique des langues

Recherche et développement

Adapting coreference resolution systems to Twitter conversations

Author:

Veronika Solopova

Supervisors:

Prof. Dr. Manfred Stede, Potsdam University

Prof. Pascal Amsili, Université Sorbonne Nouvelle Paris 3

Loïc Grobol, École Normale Supérieure

July 2020

Plan

Introduction

Starting from the 1960s, Computational Linguistics saw the rise of two new challenging tasks: coreference resolution and anaphora resolution. After 60 years, these tasks stay relevant and still show low scores (Ng 2010). With it being a complicated task by itself, there is also a lack of annotated training data for many specific linguistic genres, which makes the challenge of domain adaptation greater. In this context, Twitter conversations, which is the data type in the focus of this study, is a discourse genre especially complicated for coreference resolution, because it is highly variable and constantly changing noisy informal language, which includes abbreviations, purposeful and accidental typos, as well as Twitter-specific tokens, such as hashtags and user mentions.

As Twitter threads are dialogues with constant referencing not only to the initial post, but also the commenting users among each other, there are several exophoric pointers to the non-linguistic content in thee attached visual media, and mixed pronominal references to the same entity due to the nature of the multi-user conversations (Aktas, Berfin, Scheffler and Stede 2018).

For this study, we use several state-of-the-art resolution systems trained on OntoNotes corpus (S. Pradhan et al. 2011) as well as in domain Twitter data (Aktas, Berfin, Scheffler and Stede 2018).

First of all, we propose a statistical approach for the test/train set division, which is useful when one does not dispose of a large amount of in-domain data, and cannot afford random sampling. Then, we define the baseline experiments, comparing the benefice of in and out of domain data on the Twitter test set. Next, we consider different portions of OntoNotes, conditionally dividing them into spoken and written, to examine different training possibilities, proving that spoken OntoNotes genres, together with the Twitter training set, show better results when the current machine learning approach of choosing the quantity of data over its quality.

Furthermore, we conduct an in-depth error analysis, leading to the schema alignment process and reproduction of the previous experiments in the aligned configuration. Our best setup improves e2e-coref (K. Lee et al. 2018) resolver prediction by 21.6%. Finally, we conduct a number of experiments with normalization comparing the performance response of the two coreference resolvers in focus, and different e2e-coref models response to pre-processing.

In Section 1, we describe Twitter data specifics; in Section 2, we give detailed information of the coreference resolvers in use, as well as the reason for choosing them, with their advantages and disadvantages; in Section 3, we explain the division of train/test corpus, conduct the core experiment and present the error analysis; in Section 4, we examine the influence of the different annotation schemas in the data, measuring advantages of its alignment; in Section 5, we tackle the normalization steps, possible for this type of data and task, measuring its use for the retrained best result of Section 4 and verbal mentions excluded model, also comparing the results on two coreference resolver's 'out-of-the-box' models. The sections are followed by the discussion of the results, with prospects, and concludes.

Terminology: Coreference resolution and Anaphora resolution

Noun phrase (NP) coreference resolution or nominal coreference resolution is also known as a mention detection and chaining (Zitouni, I., Luo X. 2010). The detailed definition differs from ‘determining which NPs refer to the same real-world entity’ (Stede 2011; Grosz 1977) to (Karttunen 2020) claiming it to be an ontological question, as given an indefinite noun phrase, the classifier has to find in which case ‘there supposed to be an individual described by this noun phrase’ or ‘associate an indefinite NP with a variable and attach the binding quantifier to some sentence above the NP’, making use of all possible clues and with the slightest ambiguity possible. The task itself is also considered a search for ‘chains’ which are multiple mentions, also called discourse referents, pointing to one entity or one real-world concept in a text. As the referents are identical, they represent equivalence relation: reflexive, symmetrical, transitive (Stede 2011). As an illustration of the nominal coreference resolution, see Example 1, with all the nominal mentions referring to the same referent Boris Johnson.

Example 1: [Boris Johnson]1 is a [British [**politician]]1**, [**who]1** has served as [**Prime Minister]1** of the United Kingdom and [[**Leader]1** of the Conservative Party]1 since 2019. [**He]1** is [former[**Foreign Secretary]1]1** and [[ex-Mayor]1 of London]1.

In the context of the coreference resolution, it is essential to underline the **anaphora resolution**, as the nominal coreference resolution and anaphora resolution are closely related (Ng 2010). The scope of anaphora resolution concentrates upon identifying a nominal or pronominal antecedent for an anaphoric NP entirely dependent on this antecedent to be interpreted (Sukthanker et al. 2018), with some authors also including in its scope all the coreferential relations inside a document and the bridging anaphora (Poesio and Artstein 2008). At the same time, nominal **anaphora** resolution is usually treated as a subtask of coreference resolution, as it is considered the most frequent type and the most important for information extraction, summarization, and sentiment analysis. In contrast, the verb phrase (VP) anaphora, is less common than the nominal and usually refers to longer stretches of text, thus being harder to resolve. (Stede 2011) .

Example 2: Humans [**have increased**]2 the abundance of carbon dioxide in the atmosphere. [**This increase]2** is making big changes in our environment.

Example 2, where the verb ‘to increase’ is referential with the noun ‘increase’, can be considered an example of the VP anaphora, while sentence two of the first example, illustrates nominal anaphora, where [He] is an anaphora and [Boris Johnson] is its antecedent.

The above-described task is intuitive for the human reader, due to semantics and world knowledge we exploit, while the automatic resolvers cannot still approach our accuracy, as they implement more surface features, with real-life examples being often ambiguous. (Stede 2011)

As an example of the most ambivalent English linguistic referents challenging for the automated systems, we can use the list proposed by Sukthanker et al. 2018:

- **’One anaphora’** is an English language linguistic phenomena, manifested by ‘one’ being used as a generic referring expression;

  [One]1 can think they would have to close down.

- **Presuppositions** are indefinite pronouns such as someone, anybody, nobody, anyone used as referring expressions;

  [Nobody]1 knows [anything]2.

- **Discontinuous Sets** are the pronouns referring to more than one antecedent;

[John]1 bought a [house]2, and [Mary]3 bought a [car]4. [They]1,3 are both boasting about [them]2,4 all the time.

- **Indefinite Pronominal Anaphora** are anaphoric expressions that usually introduce entities that are new to the hearer, and need to be specified later;

  I saw **a** new [smartphone]1 today in the department store.

- **Cataphora** is an anaphoric referring expression that follows the non-anaphoric referring expression which denotes the referent the anaphoric expression points to).

  Before [she]1 came home, [Mary]1 stopped at the shop to buy some products.

- **Inferrable** or **Bridging Anaphora**, also known as an indirect anaphora, is a referring
- expression, which evokes the world knowledge of the reader or the context understanding, is presented as definite, thus building “a bridge” to a previous referent, not stating it explicitly, but as an instance apart;

  We were at that old [**castle]** today. The [**roof]** seems about to collapse, and [**the windows]** are broken.

- **Pleonastic ‘It’** is widespread in English, and it is characterized by the absence of any specified entity reference.

  It is raining cats and dogs.

- **Cleft** is a compound sentence where the copula does not serve to refer to an antecedent;

  It was Peter, **who** inspired me to draw.

- **Extraposition** is a ‘semantically empty’ pronouns, which does not point to any antecedent.

  It is possible that we will win.

Related work

Although it is largely known and accepted that the performance of coreference resolution systems (the majority of which are trained on OntoNotes) drops drastically when they are applied to unseen genres, so far as we know, the only previous study dwelling on coreference resolution adaptation to Twitter is (Aktas, Berfin, Scheffler and Stede 2018), which proves this decreased performance pattern on Twitter dialogues.

One of the few examples of domain adaptation for the coreference resolver is (Do et al., 2015), which adapts the Berkeley system (Ngoc et al. 2015) to narrative stories. In this study, the authors add linguistic features of narratives as soft constraints to the resolver, without retraining the model. The constraints here are based on the local discourse coherence, namely discourse center hypothesis (Ngoc et al. 2015), speaker-listener relations (high-lighting direct speech), and character-naming (boosting of the identification of the characters in the stories).

In contrast, Twitter-adaptation has been covered for other NLP domains, such as NER (Named Entity Recognition), as in (Ritter, Clark, and Etzioni 2011), where the study measures and compares performance using tools trained with Twitter-related and out-of-domain data. They list ungrammaticality, different vocabulary, and unreliable capitalization in Twitter as particular reasons for the systems such as OpenNLP (Apache Software Foundation 2014) and Stanford CoreNLP (Manning et al., 2016) to fail on this type of data.

Our experiments with different OntoNotes genres were mostly inspired by (Uryupina and Poesio, 2012; S. Pradhan et al., 2013) , who report varying performance in coreference resolution for distinct corpus sections. Hence, (S. Pradhan et al., 2013)'s study showed that the statistical systems in focus showed better performance on telephone conversations than on the news texts and broadcast news without assessment of the possible reasons. In contrast, in (Uryupina and Poesio, 2012) the main focus moves to 'domain-specific' and 'generic' models for fully statistical systems and those based on linguistic features (Aktaş, Scheffler, and Stede 2019).

State-of-art

The starting point for coreference resolution as we know it today dates back to the 1970s, with highly influential theoretical works by (Grosz 1977; Sidner 1979) focused on Computational theories of discourse and centering. In the 1990s, the NLP domain as a whole and coreference resolution in particular have experienced a shift from heuristic to machine learning approaches. It was also due to the appearance of the first annotated coreference corpora, published by MUC-6 (1995) and MUC-7 (1998) conferences (Van Deemter and Kibble 2000).

Another significant change was the appearance of the conferences, specifically targeting this task. It became an important topic for the general NLP conferences and also inspired interest of machine learning engineers, tackling it as a clustering task. Thus, in 1994 the world has seen the first paper on learning-based coreference resolution (Connolly, Burger, and Day 1994).

It got even more popularity with the appearance of many new Treebank corpora in many languages, mostly covering journalistic articles, the most famous of all listed by Sukthanker et al. 2018:

1) OntoNotes project (English, Chinese, Arabic) by (Hovy et al. 2006);

2) the Tübingen Treebank (German) by (Telljohann, Hinrichs, and Kübler 2004);

3) the Prague Dependency Treebank (Czech) by (Hajič et al. 2017);

4)the NAIST Text Corpus, (Japanese) by (Iida et al. 2007b);

5) the AnCora Corpus (Spanish and Catalan) by (Recasens and Martí 2010).

A new era for coreference resolution began with the CoNLL-2012 Shared-task establishing 63.37 F1 state-of-the-art on English OntoNotes. The majority of the systems since then has been trained on OntoNotes and scored with an average of MUC (Vilain et al. 1995), B^3 (Baldwin, Bagga, and Baldwin 1998) and CEAF (Luo et al. 2004) metrics, which became gold-standard for the domain (S. Pradhan et al. 2011).

Several major systems appeared at that time, including the d-coref, or the Stanford deterministic sieve system (H. Lee et al. 2011), which led to and Stanford statistical system (Clark and Manning 2015) and recent Neural System (Clark and Manning 2016b). Another example would be Berkley Coreference Resolution System (Durrett, Hall, and Klein 2013) presented for EMNLP 2013.

The most recent domain breakthrough can be associated with e2e-coref (K. Lee et al. 2017; K. Lee et al. 2018) discussed in Section 2**,** which initiated the shift from syntactic features to semantic representations. It was also one of the first works to provide end-to-end training. It also continued the trend of enhancing the system with contextual embeddings using ELMo (Peters et al. 2018). Since then, many systems based on e2e-coref saw the light, each time beating state-of-art records, the most successful of which being (Joshi, Chen, et al. 2020), (Joshi, Levy, et al. 2020) reporting 79.6 F1 using BERT (Deep Bidirectional Transformers (Kenton, Kristina, and Devlin 2017)). Another important system of 2019 is (Kantor and Globerson 2020).

Section 1. Data

Two corpora are used in this experimental study: the aforementioned OntoNotes corpus and the Twitter Conversation corpus, both possessing manual gold coreferential annotations. The differences in annotation schemas and the semi-automated alignment process are discussed in **Section 4**.

1.1 OntoNotes

OntoNotes (Pradhan et al. 2007) is a multi-genre, multi-lingual (English, Arabic, Chinese) large training corpus for Enhanced Processing used for the CONLL-2012 shared task for Modelling Unrestricted Coreference.

The OntoNotes dataset contains 2802 training documents, 343 development documents, and 348 test documents. The training documents contain, on average, 454 words and a maximum of 4009 words. We used version 4, which contains 1227K tokens in 2619 documents. A document in the CONLL-12 format, which needs a license to get access to, is an independent instance, which consists of one token per line and each sentence separated by blank lines, with blank spaces or tab-separated necessary and supplementary metadata concerning each token (Weischedel et al. 2011). The content obligatory includes document number, number of the part of the bigger document this file is part of, token number, part-of-speech tag, syntactic parse, WordNet tag, speaker information, and coreferential annotation. Number and Gender information is additional and can be added according to the shared task. The names of the files typically consist of its version_gold/auto_conll. For example, the 5th version of OntoNotes includes v4_gold_conll for train and development sets and v9_gold_conll for the official test. The 'gold' indicates that the annotation is made by hand, while 'auto' is the automatic processing output. Spoken genres include telephone conversations (tc), broadcast conversations (bc) and broadcast news (bn), while written data is represented by magazines (mz), newswire (nw), pivot text, Old Testament and New Testament, (pt)) and web blogs (wb)). The proportions can be seen in Table 1.

The three guiding principles of OntoNotes are:

• The highest inter-tagger agreement possible (for example the MUC coreference scorer gives an inter-annotator agreement of 86% (Weischedel et al. 2011));

• The biggest amount of data possible;

• Depth of representation, so that the added semantic features are as deep as possible (S. Pradhan et al. 2011).

There are some differences in the annotation scheme decisions with the previous practices. For example, the ACE program (Automatic Content Extraction research program was established by the U.S National Institute of Standards and Technology from 1999 to 2008, preceded by MUC and followed by the Text Analysis Conference):

-Attributives are not marked. (S. S. Pradhan et al. 2007) At the same time, copula "be" is annotated to transmit the attributive information.

-Adjectival modifiers like “American” in “the American embassy” (S. S. Pradhan et al. 2007) are not considered as part of the chain.

-Appositives (those having two nouns or noun phrases that refer to the same person or thing (Kelantan et al. 2018)) is annotated as a particular type of coreference. We examined OntoNotes test and train set divided for the Shared Task with quantitative analysis to calculate the average of documents, tokens, and mentions in each file to compare it to the Twitter corpus. The general statistics can be seen in Table 1, 2, 3.

Table 1. OntoNotes Data proportions

Genre	Documents	Tokens	Chains	Mentions
Bc	284	144K	4236	18K
Bn	711	172K	6138	21K
Mz	410	164K	3534	13K
Nw	622	387K	9404	34K
Pt	320	210K	6611	42K
Tc	111	81K	1931	12K
Wb	174	131K	2993	12K
Total	2632	1289K	34K	152K

Table 2. Train set averages

Average num. of docs	Average docs per file	Average tokens per file	Average mentions per file
41,333	7,63	3941,7	99,943
	The average sum of docs per file	The average sum of tokens for all files	The average sum of mentions inside all files
	192,94	133611	3255,333

Table 3. Test

Average documents number	Average docs inside per file	Average tokens per file	Average mentions per file
8.1934	4,8731	3092,79	85,4
	The average sum of docs inside files	The average sum of tokens for all files	The average sum of mentions inside all files
	49,71	23403	722,62

1.2 The Twitter corpus

The data in focus is a corpus compiled by collecting English tweets via the Twitter API, and it was constituted by Berfin Aktaş, Tatjana Scheffler, Manfred Stede (Aktas, Berfin, Scheffler and Stede 2018). It consists of 185 separate Twitter dialogues, in the forms of conversations between different users. There is no cross-document coreference.

They have a tree structure, with each tweet being a parent to the tweets it had been replied to, except for the initial tweet starting the conversation, which is a "root". This dialogue tree can be shallow, with many replies on just one level (to one particular tweet), or it can be deep when participants interact with each other across several turns. There are overall 1.756 tweets in these 185 threads, defined as a path from the root to a leaf node of a conversation tree (Aktas, Berfin, Scheffler and Stede 2018). They also report that 178 out of 278 chains possess coreference links across tweets and thus is an example of a conversational context.

The corpus has been manually annotated with coreference annotations according to OntoNotes guidelines (Weischedel et al. 2010), though with some deviations discussed in **Section 4**. The gold sentence boundaries were obtained by manually checked semi-automated segmentation. It includes Twitter ID, number of the document, number of tokens in the sentence, tokens, some of which are anonymized conforming to OntoNotes (e.d. right and left brackets into -RRB- and -LRB- accordingly) with emojis and smileys also converted into %SMILEY and %EMOJI accordingly. It also includes part-of-speech tags, and speaker information and automatic dependency parsing and Named Entities annotation from Stanford CoreNLP (Manning et al., 2016), which are used in some experiments.

Two versions of the corpus exist: with and without single mentions, also known as singletons.

"Singleton" is a cover term for mentions that are potentially coreferent, but occur only once in a document (Kübler and Zhekova 2011). See example 1 and 2 from Twitter threads:

Example 1 "With singletons annotation"

@user [Black women]1: as consistent at [the voting booth]2 as [Steph Curry]3 at [the free throw line]4.

Example 2 "Without singletons annotation"

@user Black women: as consistent at the voting booth as [Steph Curry]1 at the free-throw line.

Here we can observe the difference in the annotation. The example 1 marks all NPs (nominal phrases), such as Black woman, the voting booth, Steph Curry, the three throw line. All of them are potentially coreferent. In Example 2, only "Steph Curry" is annotated, as it is lately co-referred with Curry in the sentence "Curry never had a year that good".

See the general statistics difference in mentions in these two versions in Table 4.

OntoNotes is not annotated with singletons (Kübler and Zhekova 2011), for the purposes of evaluating the systems, which were all trained on OntoNotes, and thus can't detect them, we use the version without singletons, leaving the version with singletons for the future work.

In Table 4 and Table 1, we also provide a quantitative examination of Twitter threads to see how comparable the internal statistics are between them and OntoNotes, to be sure that it may have a considerable impact on the training when added into the training set.

Table 4. Without singletons

documents number	average tokens per file	average mentions per file without singletons	average mentions per file with singletons
185	279	13	49
docs per file	the sum of tokens for all files	All mentions without singletons	All mentions with singletons
1	48172	7053	12279

1.3 Twitter language specifics

Although some of OntoNotes genres, – namely web blogs and tv talk shows – are closely related or somehow similar to the Twitter data, it still does not include the typical challenges of our source (Aktas, Berfin, Scheffler and Stede 2018): URLs, smilies, emojis, grammatical errors and intentional and unintentional typos, prolonged vowels, mixed quotation marks, and other punctuation errors, capitalised words and even sentences, current slang, typical internet users contractions, interjections and offensive language, hashtags and user mentions handling. Consequently, it is highly problematic to get comparable to the state-of-art results on the noisy, informal Twitter conversations (Aktas, Berfin, Scheffler, and Stede 2018) with systems trained on the more standard language, like one covered in OntoNotes.

Let us consider the following examples of Twitter language specifics based on the corpus in focus:

Prolonged vowels, typos, user mentions: “**@mention Oops**, Just checked his stats .”

Emojis and slang: “@mention @mention @mention **Keyboard warrior** right there **%emoji %emoji**”.

Capitalized sentences: “/ **THAT IS WHAT EQUALITY MEANS**”.

Slang, typos, mixed quotation marks: “" **cuz** it 's not normal **an** like . . it 's been like that for ages **n** stuff ‘’”.

Links and capitalized words: “They could equally be defined as **' GREEN "** non-citizens **https://link”.**

Interjection and automatically inserted user-names: “@mention @mention @mention@**mention** @ mention @mention@ mention @mention @mention @mention @ mention @mention **Yeah lol**.” .

Section 2. Coreference resolvers

The two coreference resolvers used in this work are e2e-coref (K. Lee et al. 2017), which will be used in the retraining and normalization experiments, and the Stanford neural system (Clark and Manning 2016a), used for features creation and normalization. The decision to focus our work on e2e-coref is mainly motivated by its competitive to state-of-the-art performance with no need for extensive computational power to be retrained and full retraining of our experimental setups on GPU GeForce GTX 1080 8Gb taking from 24 to 48 hours. In comparison, the aforementioned (Joshi, Chen, et al. 2020) requires at least 32Gb of graphic card memory. Simultaneously, while the Stanford system is complicated to retrain, it is one of the quickest in prediction mode, which is useful while testing various normalization steps.

2.1 End-to-end neural coreference resolution

‘**E2e-coref’** alias ‘End-to-end neural coreference resolution’ is a system presented by Lee Kenton, Luheng He, Mike Lewis. and Luke Zettlemoyer for the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (K. Lee et al. 2017), which gained a state-of-art at that time, 68.8 F1. The task of end-to-end coreference resolution is defined by them as several decisions to assign an antecedent for every possible span in the document.

In 2018 they introduced ‘a fully differentiable approximation to higher-order inference for coreference resolution’ (K. Lee et al. 2018). The renewed approach uses the ‘antecedent distribution from a span-ranking architecture as an attention mechanism to iteratively refine span representations’ (K. Lee et al. 2018) so that the model could consider multiple ‘hops’ in the predicted clusters. They also introduced a coarse-to-fine approach using a less accurate but more efficient **‘bilinear factor’,** enabling more aggressive pruning without hurting accuracy. In machine learning, ‘pruning’ is a technique used in search algorithms to reduce the complexity of the search space. It also improves predictive accuracy, decreases overfitting and complexity of the classifier.

Augmented with both contextual word embeddings (ELMo, Peters et al. 2018)â and hyperparameter tuning, the final approach achieved 73.0 F1, setting a state-of-art.

Their model consists of two important steps to learn a conditional probability distribution to produce the correct clusters:

1)Computing spans embedding (in Natural language processing an ‘embedding’ is the generic name for a number of language modeling and feature learning techniques, where vocabulary entities are mapped to vectors of real numbers) and a “mention score”. Low-scoring spans are pruned out, and only the highest-scoring ones are considered as candidates for referential entities.

2)The antecedents are scored from pairs of span representations. The final coreference score of a pair of spans 'is computed by summing the mention scores of both spans and their pairwise antecedent score’(K. Lee et al. 2018).

The authors claim that the most crucial information transmitting coreference links is ‘the context surrounding the mentioned span and the internal structure within the span’ (K. Lee et al. 2017), which they encode using bidirectional LSTM (Hochreiter and Schmidhuber 1997). Normalized as a unit vector, the word embeddings is a concatenation of 300-dimensional GloVe embeddings (Pennington, Socher, and Manning 2014) and 50-dimensional embeddings from Turian et al. (Turian, Ratinov, and Bengio 2010). Out-of-vocabulary words are represented by zeros.

Speaker information is encoded as a binary feature, which shows if a pair of spans share the same speaker. All metadata features (speaker, genre, span distance, mention width) are represented as learned 20-dimensional embeddings. The performance degrades by 1.4 F1 without this data (K. Lee et al. 2017).

The hidden states in the LSTMs have 200 dimensions. Each feed-forward neural network consists of two hidden layers with 150 dimensions and rectified linear units as in (Nair and E. Hinton 2010). The Adam Stochastic Optimization algorithm (Kingma and Ba 2015) is used for learning with a minibatch size of 1 (document). The model is trained for up to 150 epochs or 400 steps with early stopping based on the development set. Code is implemented in Tensorflow GPU version 1.13.1, with many features deprecated in the current Tensorflow 2.

2.2 The Neural Stanford System

The Neural Stanford System was proposed by Christopher D. Manning and Kevin Clark from Stanford University in 2016 (Clark and Manning 2016b) and improved in (Clark and Manning 2016a), marking the definite transition from the previous deterministic (H. Lee et al. 2011), (Raghunathan et al. 2010) and statistical (Clark and Manning 2015). It uses a modified reward-rescaled max-margin objective, initially proposed by Wiseman et al. (Wiseman et al. 2015), which resulted in considerable improvements over the previous state-of-the-art on the English and Chinese portions of the CoNLL 2012 Shared Task data with 65.73 F1.

This model is a neural mention-ranking model, which scores pairs of mentions for their likelihood of coreference. The following approach was widespread at that time, as it is fast, scalable, and simple to train (Clark and Manning 2016a). Taking as an input a mention and candidate antecedent, the mention-ranking model produces a score for the pair indicating their compatibility for coreference with a feed-forward neural network (Clark and Manning 2016a). The input layer gets for each mention various words and groups of words depending on their syntactic role, represented by embeddings and the average of the vectors of each word in the group respectively, with a small number of additional features used, including distance in tokens, string matching, and speaker identification features (Clark and Manning 2016b). The input gets passed through three hidden layers of rectified linear (ReLU) units (Nair and E. Hinton 2010). During the test, the mention-ranking model links each mention with its highest-scoring candidate antecedent, and the system prioritizes them (Clark and Manning 2016a).

The system is part of the CoreNLP toolkit (Manning et al., n.d.), and uses the annotations produced from this system: Part-of-Speech, Dependency parses, Named entities.

We tested different input options of the Stanford system: raw text, XML, and CoNLL formats. There are several differences between the possibilities they present: we found out that there is no option to introduce speaker information using txt and XML formats, for instance.

Section 3. Baseline experiments

3.1 Reproducing the results

The starting point of our reasoning was the idea that OntoNotes include several genres that can have a negative influence on the prediction of the Twitter test documents. After closely reproducing the results reported in (K. Lee et al. 2018) and getting similar results with a newly retrained model and the one available in the e2e-coref package, we also measured if Twitter introduces any confusion into the model if added into the training set. We report that adding Twitter corpus into OntoNotes training set improves the performance on the official OntoNotes test by 0,83% compared to (K. Lee et al. 2018). We additionally led several tests measuring our latest changes to the data, anonymizing smilies and brackets, and sentence boundaries revision, which brings a 1% gain to the performance on the Twitter data test. Finally, we measured the influence of singletons in the test and train sets getting substantial performance loss. The results can be seen in Table 5.

Table 5 Preparatory tests

Experiment	F1 overall
Twitter data full tested with Lee Kenton’s package model	45,79
Twitter data full tested with the retrained model	44,68
OntoNotes official test with the retrained model	73,84/ 73,4 (package model)
full OntoNotes with 39 Twitter documents in the train, tested with 9 Twitter documents +full OntoNotes test	73,64
full Twitter in training tested on OntoNotes only	**74,23**
package model tested with Twitter before changes	44,29/45,2 (package model)
Test on the full Twitter with singletons with a retrained model with OntoNotes	33,31
39 with singletons in train+ontonotes/test on 9 Twitter documents with singletons+ OntoNotes test.	71,16

3.2 Test set composition

To see how different systems would perform if we enrich OntoNotes training data with Twitter threads, and how specifically we can use it in order to increase performance on this type of data, we divided our data into a training set and a test set. As we do not dispose of a large number of Twitter examples, we decided to make the 10-15% of the total threads the most representative possible, so that, on the one hand, threads with a little number of annotations do not enter the test set, and on the other hand, those we chose, to have a number of Twitter language features, that notably differ it from OntoNotes and thus induces errors.

In order to achieve informative results, as the data is not linearly distributed and highly variable, we selected a representative test set not via random sampling, but through statistical analysis of three features:

1) Number of tokens per document (which also implicitly shows the length of the thread);

2) The total number of chains per document;

3) The total number of mentions per document.

Table 7. Example of a test set selection

Document id	**Tokens**	**Mentions**	**Chains**
0	47	4	1
14	166	23	6
20	773	177	43
109	105	29	20

To faithfully represent threads of all lengths, we determined the documents where these variables are situated either on the median or in the first and fourth quartiles of the respective distribution while omitting apparent outliers. Because of the linear correlation of the three parameters, we could make sure only to select the documents where all three are in the same range of their distributions. In Table 6, we show document 0,14 and 20, where all three features lay on the respective distribution quartile (median, first and fourth quartile respectively), and document 109, where tokens lay in the third quartile, mentions in the first and chains in the fourth. Hence, document 109, and those alike were left out.

Among the pre-screened files, we checked each document, marking features of the annotated mentions (person, number, gender), and Twitter phenomena (hashtags, user names, pronouns with typos, etc.). With this information, we excluded threads without enough coverage and variability of the phenomena in focus. One of the gains from accepting the long threads of the fourth quartile was also the coverage of the poorly represented second person plural and third person plural by the short and medium-size threads.

As the threads are not evenly distributed in their total length, we compared the average, median, and sum for each of the three characteristics in the whole corpus with those of the determined test set, confirming that all values lie under the 15% threshold of the total number. You can see the final distribution in Table 8.

Table 8. Final test/train set distribution

	Documents	Tokens	Chains	Mentions
Train	165	44912	1596	6589
Test	20	3260	134	462
All	185	48172	1730	7053

3.3. Train set experiments

With the first test (Test A), we measured the performance we get on the previously determined test set with the model trained solely on OntoNotes, getting 39,77 F1.

We further proceeded to measure how the test set responds to the different proportions of Twitter data to OntoNotes in training set as well as how different genres influence the result, being especially interested in the spoken genres against written genres comparison, as several studies prove that Twitter bears many features of spoken – rather than written – language.

Firstly, with Test B, we measured the performance of the whole OntoNotes train set with the remaining 44912 tokens and 165 documents of the Twitter corpus, which formed roughly 3,5% of the train set for this test. The results show a considerable increase in performance: up to 20% comparing to using only OntoNotes training set..

We distinguished spoken, spontaneous language from written or edited texts choosing tc (telephone conversations) and bc (broadcast conversations) to represent the first and mz (magazines) and nw (newswire) for the latter.

Further, Test C included spoken genres with wb (web blogs), as the only internet genre available in OntoNotes. It also presupposed a considerate augmentation of the Twitter proportion in the training set (from 3,55 % to 11%), and spoken genres (from 17,6% to 64%). A slight increase in overall performance is less than one percent (0,81%).

In Test D, we excluded web blogs, and training only with spoken genres: broadcasts conversations,and telephone conversations (bc, tc respectively), representing 83% of the set, with Twitter training portion being almost 17%. This augmentation led to a considerable improvement of almost 5% (4,71%) in comparison to Test B, with an overall score of 64,27% F1.

Having a doubt whether the improvement was caused by the full filtering of written genres, which could be noise for the training set for this task, or an increase of the Twitter against over types of data, we decided to proceed to Test E, which seeks to reproduce the same proportions of Twitter against over types of data (17%), changing spoken data to written genres documents newswires and newsgroups (nw and mz). Test E scores overall F1 62,39%, which is roughly a 1,88% loss with respect to Test D, which we consider confusing and controversial results, which can be explained by the quality of documents more than genre suitability. Thus, we proceeded to Test F, which was trained only using the Twitter training portion. We are aware that this model cannot be considered robust due to the small amount of data used so that it will not be general enough to handle the changes inTwitter language trends even in a year. However, the test is to show if the OntoNotes presence in the training set can be rather a negative factor in general. Notwithstanding, Test F, with 61,86 % F1, is 2,5% worse than Test D (with spoken genres) and only 0,53% worse than Test E with written genres. Again the difference lies in the small margin, but spoken genres setting prove to be better. At the same time, Test F proves that using the whole OntoNotes corpus brings confusion to the model, as without the other genres, the system gains at least 2 points.

Finally, we examined the last two scenarios, increasing the number of documents in spoken and written genres by the same amount (Test G and Test H respectively). As we already exhausted the capacity of spoken genres of the official Shared Task 2012 training set version, we managed to augment it by collecting all the documents from the official development and test set, as well as spoken genre documents included into the newer versions of OntoNotes (Release 5.0) (Weischedel et al., 2010), resulting in 91420 new tokens, 13031 new annotations, and 4712 chains, altogether being 40% increase. The same 40% increase is preserved for the Test H, with new documents added from web blogs and full newswire genre, only taken from Official training set.

Hence, for both Test G and Test H, the proportion is 12,45% of Twitter against 87,55% of OntoNotes of the particular genre, which leads to a decrease of the Twitter proportion of 5,15%, in contrast to Tests D, E. In terms of performance, Test G, with spoken genres outperforms Test H with written genres by 4,2%, with F1 being 65,6% and 61,41% respectively, therefore overall proving that spoken OntoNotes genres are beneficial to the task of coreference resolution in the Twitter threads context, and giving us a new baseline of 65,6% for the further experiments. The results and Twitter proportions are summarised in Table 9, while the exact numbers of tokens, mentions, and chains per genre in each experiment can be seen in Appendix 1.

Table 9. Baseline experiments

Test	Twitter % in tokens	Twitter % in annotations	F1
A (OntoNotes only)	-	-	39,77
B (Full OntoNotes, Twitter)	3,55	4	59,5633
C (bc, tc, wb, Twitter)	11	13	60,37
D (bc, tc, Twitter)	16,9	17,6	64,27
E (mz, 1/2 nw, twitter)	16,8	19	62,39
F (Twitter only)	100	100	61,86
G (Twitter; bc, tc, augmented by 40 %)	12,45	13,65	65,6
H (Twitter; mz, nw, wb, augmented by 40 %)	12,45	15,6	61,41

Looking further into experiments, we also measured the statistics concerning the prediction of the first and second person singular and plural, which usually does not require difficult reasoning for the coreference resolution system, against third person singular and plural full of ambiguous real-world examples. We excluded from all the predicted test files all the first and second person annotations and measured the difference between them and the systems output, to see the exact number of predicted mentions and chains in the different settings (reported in Table 10), except for the Test A, which does not include Twitter corpus, and Test C (as weblogs mixed with spoken genres does not satisfy our newly defined interest in spoken/written genres comparison). The numbers do not reflect the correct assertions, but the exact number in the predicted file.

We can see that all the models except for the Test D and E overproduce first and second person annotations, especially Test G, trained only with Twitter, producing almost two times more annotations, when in the golden standard. As for the third person annotations, neither of the models reaches even the approximate amount set by the golden variant, but Tests D and E have the nearest number (390, 389 respectively) against 502 in the golden version.

Table 10. First-second person/third person performance

	Annotations	Chains	First-scond person	third person	First-second chains	third person chains	F1 full	F 1 without first-second person	Difference
golden	546	159	44	502	16	143	100	95,27	-4,73
onto, tw Test B	430	135	75	355	21	114	59,5633	52,06	-7,5
Tc,bc,tw Test D	428	133	38	390	13	120	64,27	59,45	-4,82
Nw,mz, tw Test E	433	125	44	389	16	109	62, 39	58,15	-4,24
twitter Test G	408	116	81	327	18	98	61,86	52,34	-9,52
Augm. spoken Test F	429	132	72	357	21	111	65,6	57,38	-8,22
Augm. written Test H	429	120	70	359	19	101	61,41	53,37	-8,04

As for the predicted chains, the amount of first-second person annotations is also logically generally overproduced, with test D being the only underproducing and Test E, the only one matching the gold standard. At the same time, third person chains are underproduced, with the worst results for Test G, and the best for Test D.

Considering the last two tests with augmented portions of spoken and written genres, it is worth underlining the similarity of the numbers of overall annotations and first-second person annotations. However, the underproduction of the third person chains in the written genre experiment probably leads to the reported 4,2% loss.

Comparing the evaluation metrics of the output to the gold version and predicted version both with first-second person annotations excluded, we compared the losses of F1 in two versions. The “normal “loss we get, comparing the gold version to the gold version with first and second person excluded, is around 5%. The indicator is only satisfied by Test D and E, proving them to be the best performing from the third person correct annotations. However, all the other tests lose from 7,5% up to 9,5% for Test G with first-second person exclusion, which means that they were better predicted in these terms and balanced the problems in the third person predictions.

Mainly, this is interesting in terms of Test F, which outperforms all the other tests in F1, but has 8,22% loss in this experiment. If we compare it to the closest competitor Test D with 64,27%, we can see that the later is much more balanced in all senses, by the number of annotations, and the correctness of third person predictions. At the same time, it means that by overproducing first-second person annotations and covering most of them right, it gains these additional percent. However, Test D can still prove to be useful in future tests.

3.4 Error analysis

In order to determine which preprocessing or post-processing steps can be undertaken, we perform a comparison of the errors of the system trained solely on OntoNotes (as in K. Lee et al. 2018)â on the predicted test file , i.e.the output of Test A, and the best result we got so far with Test F.

We retrieved in-depth statistics concerning the gold annotation in the test set. This time we took into account not just annotations, which in previous experiments were defined in a simplistic way, as any token with annotation, which leads to counting multi-token mentions several times, but as mentions in their multi-token borders. Although the initial approach is not fatal for the surface statistics needed to evaluate proportions of data for the previous tests, the error analysis needs some more exact and non-ambiguous numbers to compare. Besides using Annalena Kohnerts program, we measured the exact number of first, second and third person pronouns, of non-standard pronouns (e.g. "˜ur, "˜im), @-mention which are part of syntactic structure and automatically inserted @-mentions, which are, nevertheless, mentions of a coreferential chain, as well as mentions fully being a hash-tag and multi-token mentions that include hashtags. As for the chains, the statistics include the number of chains, the averagemention length in tokens, the number of chains with pronouns, and of chains with non-standard pronouns (see Table 11).

Comparing the results from Test A, Test B, and Test G, we notice that both mention identification, conferential links, and non-coreference links performance influenced the overall F1 increase, reflected in the results. Test A copes relatively well with identification of mentions, scoring F1: 55.28%, but does not score far better than a random choice for coreference and non-coreference links (22,96%, 24,7%, respectively). With the addition of the Twitter training portion to the OntoNotes train set, we see an immediate increase of almost 20% fromention identification, but also a 35-37% increase in coreference and non-coreference links (Identification of Mentions: F1: 74.49%; Coreference links: F1: 59.8%; Non-coreference links: F1: 59.54%). Finally, the best proportion between Twitter and spoken genres in the training set actually degraded the identification of non-coreference links, while only slightly increasing that of coreferential links, and thus the better part of the 6% gain we report from Test B to Test G, is mostly due to the better identification of mentions (Identification of

Mentions: F1: 79.61%; Coreference links: F1: 61.03%; Non-coreference links: F1: 57.43%). This empiric evidence comforts the hypothesis that the most challenging part for the systems trained on OntoNotes is previously unseen Twitter-specific types of mentions.

3.4.1 Mention Identification

Table 11. Test corpus Golden Mentions and Chains Statistics

All gold mentions	
Number mentions	**462**
Mean mention length (in tokens)	**1.673160173160173**
Number Pronouns	**233**
first Person Pronouns	**59**
second Person Pronouns	**69**
third Person Pronouns	**70**
Number non-standard Pronouns	**7**
@-mentions	**8**
Tweet-initial @-mentions	**1**
Hashtag-mentions	**11**
Mentions including hash-tag	**13**

All Gold Chains	
Number Chains	**134**
Mean chain length	**3,44776119402985**
Mean mention length (in tokens)	**1.762984254551419**
Number pronoun chains	**24**
Chains with non-standard pronouns	**7**

We consider all the possible cases for predicted mentions: they are consistent with gold, the gold annotation is missed, and an extra mention is predicted, which is not annotated in the gold standard. Test A although predicts 72 mentions less, produces 48 and 46 more @-mentions, Tweet-initial @-mentions, seriously underproducing first and second person pronouns in comparison to Test G and Gold Standard. At the same time, the average mention length is closer to the reference in Test A, while Test G, on average, produces shorter mentions when whose of the gold version.

At the same time, we see a serious increase in the correctly found gold mentions 212/334, 45,8% versus 72% correct at the current state. There is a spike in pronoun resolution with 215 out of 233 being correctly resolved in Test G, and only 150 in Test A. At the same time, as we stated before, third person pronouns have not seen considerable changes between the two experiments, with most points gained in first-second person pronouns.

Concerning the @-mentions, in Test A, of the 51 user-names it predicted as mentions, only five were correct, while in Test G the three detected ones were correct, while other five were false negatives. It is still a very positive change, which brings improvements in both recall and precision, as we produce less false positives, and all the annotations are true positive. We also note a significant gain in hashtags detection, which were completely ignored by Test A, while Test G correctly identifies every other entity of this type.

It is also worth mentioning that e2e-coref by itself with the "out-of-the-box" model is not bad at identifying non-standard pronouns. Indeed, in Test G, the system finds all 100% of them.

As for the extra predicted mentions, in Test G, they are two times fewer (93/43), generally not only due to the more balanced @-mention annotation production but also due to the third person pronouns. At the same time, overproduced first-second person pronouns were slightly higher in number than in Test A. Test G also overproduced only two mentions with hashtags. The exact numbers for mention identification can be found in Appendix 2.

3.4.2 Chain Identification

In terms of number of predictions, we can see that there is not much statistical difference in quantity, but there are qualitative differences between the tests. Test G shows three times more correct chains, which is still only 30% of all the gold chains, and the chains produced are a bit longer than in Test D. It is at the same time better at pronoun chains, but has on the other hand not found any full chain with non-standard pronouns. Analyzing the wrong chains from Test A, we also see that a lot of correctly found gold mentions are linked to incorrect NPs, from 88 to 58 in Test G, with six of them containing non-standard pronouns. However, more of the incorrectly linked mentions in Test G are now pronouns, with a massive increase in mean chain length for the completely predicted gold chains, which possess extra mentions. At the same time, we can explain this in terms of the general increase in predicting correct and nearly correct chains than in Test A, which can be seen as an improvement

Finally, the most crucial error was found while looking into alternative spans (Table 12), as we notice many overlaps.This suggests that the system predicted the test set using a different annotation schema than the one in the Twitter gold annotations. Thus, before experimenting further, we decided to look into the annotation differences between the guidelines of OntoNotes those used for the Twitter corpus to find potential clues to align them semi-automatically

Table 12. Alternative spans

Error type	Test A	Test G
Alternative span beginnings	8	7
Alternative span ends	12	14
Intermediate spans	0	4
Span overlaps at the beginning	0	0
Span overlaps at end	0	1
Span with pre-context	8	5
Span with post-context	11	11

Section 4. Corpus alignment

One of the important sources of errors turned out to be the differences between the choices made for the Twitter corpus annotation (Aktas, Berfin, Scheffler and Stede 2018)â and OntoNotes annotation guidelines (Weischedel et al. 2010) on several points, which on a closer look on the Test Gs output file brought much confusion to the system.

So, we defined several linguistic phenomena, that are not treated the same way in two data sources:

Copula

Copulas are finite verbs that lose their lexical sense and serve a fully grammatical purpose of linking the subject of a sentence to a subject complement. Expressions using copula were annotated in the Twitter corpus as coreferential, while in OntoNotes guidelines, it was explicitly mentioned that they must not be annotated. Coreference Guidelines for English OntoNotes - LDC Catalog states:

"Relationships signaled by copular structures will be captured through word sense tagging, and annotators should not mark coreference between the two elements: (37) [John] is [a linguist] (no co-ref.)"

Twitter: [Todd]1 is [a very loyal guy]1

OntoNotes rules: [Todd]1 is a very loyal guy

Possessive markers

In particular, we consider the genitive mark "s". While it is not included in the mention boundaries in the Twitter annotations, it is for the OntoNotes annotations.

Twitter:[@Borisjohnson]1 s speech

OntoNotes rules: [@Borisjohnson s]1 speech

In the output file, all such occurrences (both copulas and possessives) were predicted according to the Twitter corpus convention and thus did not impair the results, but we think that any future new data can nevertheless get this error, so it is worth correcting.

3. Reflexives

Reflexive pronouns are used when the subject and the object of a sentence are the same. They are inherently anaphoric, but they become ambiguous when they are situated right after the antecedent noun, e.d "˜you yourself as OntoNotes guidelines claim it to be one unique mention. At the same time, in Twitter corpus they are annotated as separate coreferring mentions:

Twitter: [The society]1 [itself]1

OntoNotes rules: [The society itself]1

4. Appositives

An appositive is a noun or NP, which follows another NP, providing further information about the first one. In the Twitter corpus, appositions are annotated separately from the noun they corefer with. In OntoNotes, appositions are merged with the NPs they define. While aligning these schemas with the OntoNotes guideline, appositives were the primary cause of the decrease in the number of chains , as the merging procedure creates singletons, which had to be removed from the corpus as a post-processing procedure.

Twitter: e.g. [you]1 [guys]1

OntoNotes rules: [you guys]1

5. Generic nouns and pronouns.

In OntoNotes, no generic noun or pronoun is annotated, while it is not the case for the Twitter corpus. However, we considered that the OntoNotes guidelines and the Twitter corpus annotations followed different definitions of generality and thus are hard to align, except for generic "you" instances. Hence, we only removed these chains from the Twitter corpus.

Twitter: Are [you]1 prepared for the coming of Jesus?

OntoNotes rules: Are you prepared for the coming of Jesus?

6. Relative clauses

6.1 Headless.

"A relative clause which apparently lacks a head is called a free relative clause, also sometimes called a headless relative (though some argue that the head is present syntactically but phonologically empty, and hence that this is a misleading term)."(Comrie, Asher, and Simpson 1995)â

Twitter: [What]1 [you]2 say is [true]1

OntoNotes rules: What [you]1 say is true

6.2 Unrestricted

The relative clauses that serve as an attribute to the noun they "˜explain.

Twitter: The same [logic]1 [that you apply to other]1

OntoNotes rules: The same [logic that you apply to other]1

7.Verb mentions

While Twitter corpus is annotated only with nominal coreference, in OntoNotes, verb mentions are annotated if they corefer with a nominal mention (S. S. Pradhan et al. 2007)â . For the sake of comparability with other systems trained on the OntoNotes data, we decided to perform two separate sets of experiments, described in Section 4.3 and Section 4.4.4: one keeping verb mentions in OntoNotes and the second, automatically removing them. Excluding verb mentions also leads to a decrease in the number of chains, also due to the creation of singleton mentions.

Twitter: Humans have increased the abundance of carbon dioxide in the atmosphere. This increase is making big changes in our environment.

OntoNotes rules: Humans [have increased]1 the abundance of carbon dioxide in the atmosphere. [This increase]1 is making big changes in our environment.

4.2 Alignment results

As a result of the alignment process, 11% of the chains (1728 vs. 1534) and 10% of the mentions (7067 vs. 6354) were eliminated in the Twitter corpus. For the test set, it is 8% of chains and 14% of mentions, and for the train set, it is 11% of chains and 9% of mentions. As we can see, the number of mentions excluded from the test set is slightly higher than the overall statistics, which means that a smaller amount of mentions to predict can influence the evaluation outcome.

Table 13. Test/train set distribution before and after alignment

	Chains	Mentions	Chains mod	Mentions mod	Diff chains	Diff mentions
Train	1596	6589	1411	5946	185	634
Test	134	462	123	408	11	65
All	1730	7053	1534	6354	196	699

4.3 Aligned experiments

We decided not to proceed with the previous Test setting C (which included data from web blogs and spoken genres, because as we mentioned above it did not satisfy our newly directed interest towards written/spoken genres comparison. We also reorganized the order of the experiments. Hence, we first measured how a model trained on OntoNotes performs on the aligned Twitter test set (Test A). We report a 6% increase in performance compared to the experiments with the unaligned data with 45.18 F1 (see Table 14), which is still almost 28% lower than the result reported on the official OntoNotes test set. As a second baseline experiment, we chose the training with solely Twitter data, with a 1% loss (60.8 F) which, as we mentioned before, can be explained by the decrease of the number of mentions to predict. Still, it improves significantly on Test A and highlights the difference between in-domain and out-domain training.

For Test C, we chose to train with all of the OntoNotes and Twitter data, scoring 62.51 % F1, which shows a 3% gain from alignment, 17% in comparison to aligned Test A, and a 2% gain over Test B.

The aligned recreation of the experiment with spoken genres and twitter (Test D) achieved 66.8%, our best result so far, increasing our first result for almost 27%, while Test E, with written genres and

Twitter, repeated the pattern, showing a 5,5% loss, while also losing 1% in comparison to the unaligned test.

Finally, the augmented experiments, which were the most successful with unaligned annotations, did not prove to be the best performing this time, with augmented spoken losing 1.5% and augmented written losing only 0.2%, showing no statistically significant difference with not augmented Test E. While Test G (augmented written genres) can be explained by the scarcity of the written data, and the fact that by augmenting the number of tokens we only slightly augment the number of annotated mentions, which leads to no change in the result. However, the spoken augmented setting still gives confusing and ambivalent results, inspiring further work in the qualitative evaluations. At the same time, we consider Test D a highly positive result, as it proves that the standard machine learning approach of giving more data to the system, without qualitative assurance, loses to the correctly chosen proportion of data, qualitatively closer to the domain in focus, and therefore overall proving that the spoken OntoNotes genres are beneficial to the task of coreference resolution in the Twitter threads context. The data setup statistics can be seen in Table 15.

Table 14. Aligned results

Test	Recall	Precision	F1	F1$^{1 \text{ (not aligned)}}$
A - onto	39.29	53.89	45.18	39.77
B - tw only	53.15	71.02	60.8	61.86
C - onto + tw	57.76	68.25	62.51	59.56
D- spoken+tw	**60.72**	**74.39**	**66.8**	64.27
E - written+tw	55.98	67.7	61.25	62.39
F-augmented D	57.56	72.39	64.13	**65.6**
G- augmented E	57.47	65.67	61.29	61.41

Table 15. Experimental setup

Experiments	Tokens	Chains	Mentions
A - onto	1223K	85K	197K
B - tw only	45K	1.5K	6.3K
C - onto + tw	1268K	86.5K	203.34K
D- spoken+tw	269K	8K	37K
E - written+tw	269K	6.5K	30K
F-augmented D	360K	10K	47.5K
G- augmented E	360K	8K	32K

4.4 Error analysis

4.4.1 Pronouns

To make sure that the experiments improvement pattern is not exclusively caused by correct predictions of the pronouns, we repeated the measures of the performance of the aforementioned experiments, excluding 1 and 2 persons singular and plural (see Table 16). Our initial measurements show that the grammatical person distribution of the pronoun gold mentions is even in the dataset. However, more third person pronouns are resolved than first and second person. In Test A, while in Test B with Twitter training data, which has more pronoun instances, the performances improve only for the first and second person, not showing changes for the third person. In Test D, overall pronoun performance is slightly worse (0.905) because, in Test B and D, all the data is represented by conversions, which naturally includes many first-second person pronouns.

From Table 16, we can see that all the later tests improve over the baselines A and B, which means that the gains are not only due to deictic pronouns but also to the detection of other anaphoric expressions.

Table 16. Third person prediction performance

Experiments	Recall	Recall full	Precision	Precision full	F1	F1 full
A - onto	37.39	39.29	50.21	53.89	42.6	45.18
B - tw only	47.27	53.15	66.58	71.02	55.27	60.8
C - onto + tw	46.9	57.76	65.43	68.25	54.67	62.5
D - spoken+tw	49.43	60.72	71.69	74.39	58.3	66.8
E - written+tw	50.01	55.98	64.56	67.7	56.32	61.25
F - augmented D	46.32	57.56	68.21	72.4	55.11	64.13
G - augmented E	51.9	57.47	63.72	65.67	57.2	61.29

We further focus the comparison on the best performing Test D and the baselines Test A and B.

Table 17. Predicted mentions statistics

	Gold	A	B	D
Pred. Mentions				
All	408	305	307	334
Usernames	8	51	6	5
Tweet-initial	1	44	0	0
Hashtags	11	0	4	5
Correctly predicted				
All	408	218	143	293
Average mention mean length	1.64	1.41	1.13	1.18
Pronouns	219	149	199	194
first	57	38	53	50
second	64	26	63	62
third	68	60	61	59
Usernames	8	6	5	5
Tweet-initial	1	1	0	0
Hashtags	11	0	3	5

4.4.2 Mention identification

For all the tests, the average token length of mentions additionally predicted by the system is significantly longer (0.05) than the one in the golden annotation. The higher the proportion of OntoNotes (whose mentions are on average 0.72 tokens shorter than in Twitter) in the training data, the longer those predictions are. Hence, we can report a tendency to predict longer spans (especially when training on OntoNotes).

Speaking about Twitter-specific mentions, hashtags and usernames caused many errors in Test A, with the out-of-the-box system. In the replying tweets, user mentions, in the beginning, are usually automatically inserted, and consequently are not part of the syntax, and as such are not considered markable in Twitter gold annotations. Thus, the model (A) trained without these features in the training set, mispredicts these as mentions, entirely ignoring the hashtags. The addition of the Twitter data in Test B instantly improves the performance in many ways: Twitter-specific mentions detection, ignoring tweet-initial usernames, predicting almost all of the hashtags, with those predicted being correct. Test D further improves on hashtags, which participate in syntax, with no visible changes in usernames and a slight decrease in pronouns detection.

Regarding verb mentions, which were not aligned in this setup, four predicted verb mentions of which two are correctly linked with the demonstrative pronoun "that", are counted as erroneous predictions, and thus interfere with the evaluation. After adding Twitter data in the training data, no verbal mentions are predicted (Test D). Which is why we conducted a series of tests, aligning verb annotation in Twitter and OntoNotes, which are described in **4.4.4**

4.4.3 Chain prediction

Test B improves the number of correctly predicted chains compared to Test A while producing 20% fewer chains. This number further improves in Test D, almost doubling the Test A numbers, while predicting almost the same amount of chains. At the same time with each test, more partially correct chains (with alternative beginning and endings) are present. The number of completely missed entities is reduced by 51.3%, with chains consisting only of identical strings profiting the most from the spoken genres training set in D.

4.4.4 Verbal mentions alignment and nominal coreference resolution

Table 18. Nominal coreference results

Test	F1	F1 [2]
A - onto	45.18	50.99
B - tw only	60.8	-
C - onto + tw	62.51	**65.0**
D- spoken+tw	**66.8**	63.76
E - written+tw	61.25	64.60

As the last experimental setup in this section, we repeated the same tests automatically, excluding verb mentions and singletons appearing after these procedures. The decision to make changes in the OntoNotes and not Twitter corpus is motivated by the fact that the exclusion could be made automatically while annotating Twitter corpus with verbs, which would be extremely time-consuming, even though it is a considerably small corpus. At the same time, this configuration lets us examine purely nominal coreference resolution, which is supposed to be a less confusing task. The process reduced OntoNotes mentions by 2.4% and chains by 3.6%

We only decided to repeat training for four setups, excluding augmented spoken and written tests, and obviously "only Twitter training test". Surprisingly, for the nominal coreference resolution, the highest performance is achieved with test C, with full OntoNotes and Twitter training data, while written genres training also slightly outperformed spoken-only training. These variations motivate looking further into the specific effects of different training data combinations and how verb annotations have influenced the nominal coreference resolution task. The comparison of these results can be seen in Table 18.

To look deeper into the causes of this departure from the pattern seen in the previous experiments, we conducted a quantitative and a qualitative analysis of the OntoNotes genres in terms of verb mentions and chains becoming singletons and thus excluded with them. We also compared the portions included in spoken and written genre experiments.

Thus, from Table 19, we can see that in absolute numbers, pivot texts (pt) are the ones most influenced by the verb exclusion, but we did not use them in our genre-specific experiments. However, considering that all the genres are highly different in the number of tokens, the relative statistics are informative. According to them, while the bn (broadcast news) genre lost the biggest number of verb mentions (3.3 % of all), bn and bc (broadcast news and conversations) lost the most mentions, considering both verbs and other mentions, what were excluded as new appearing singletons. The least influenced genre was "magazines", also having the smallest amount of verbal mentions in general.

Considering the experimental setups we tested before, in relative numbers, it was the spoken genres experiment that proportionally to its size lost the biggest amount of all annotations it had (3%). Hence, of all 3705 mentions excluded from the OntoNotes by this procedure, 25% corresponded to spoken genres. Thus, visibly the exclusion of the verbs had a negative influence on training, especially with spoken genres, because of the immense loss of the training material.

We also looked into the nominal mentions eliminated by this procedure. As it could be expected, the most frequently occurring ones are demonstrative/relative pronouns "this", "that" and the personal pronoun "it". As for the verbs, the most frequent are the auxiliary verbs "be" and "have", as well as "say" in different tenses (see Table 20). However, there are less frequent instances, as verbs "kill", "meet", “attack", "fire", "blow", "report", "name", "run", "get", "suffer", "tell", and "give". These verbs are usually in the form of past participle two or three.

Table 19. Verb mention statistics per genre

Genres	Verb mentions	Mentions total/verb	Chains before	Mentions before	Chains now	Mentions now	Diff chains	Diff mentions	Other	Excluded to total
bc	435	2.3404	4236	18586	4065	17980	180	606	171	**3.3**
bn	513	**2.4061**	6138	21320	5948	20626	190	694	181	**3.3**
mz	150	1.1135	3534	13471	3449	13236	85	235	85	1.8
nw	**622**	1.8077	9404	34408	9189	33584	215	**824**	202	2.4
pt	380	0.9029	6611	42086	6291	41386	**320**	700	**320**	1.6
tc	207	1.7046	1931	12143	1802	11807	129	336	129	2.8
wb	172	1.3954	2993	12326	2855	12016	138	310	138	2.5
Full onto	2479	**1.4**	34847	154340	33599	150635	1257	3705	1226	**2.4**
Written	193	**1.1086**	4553	17408	4444	17108	109	300	107	**1.7**
Spoken	**642**	**2.0892**	6167	30729	5867	29787	309	**942**	**300**	**3.0**

Table 20. Appearing singletons > 5 times

Appearing singletons								
Tokens	Bc	Bn	Mz	Nw	Pt	Tc	Wb	All
That	**113**	**119**	11	**89**	105	**93**	**58**	588
This	17	17	**56**	67	**177**	3	54	391
It	41	48	18	44	38	33	26	248
Other tokens	0	5	0	6	0	0	1	12
Verbs								
Is/‘s	29	10	10	14	13	16	12	104
Say	7	18	0	12	0	0	0	37
Are	9	0	0	0	20	6	0	35
Have	7	0	0	0	8	18	0	33
Be	7	7	0	7	0	0	0	21
Other tokens	31	48	0	32	21	14	0	146

After consideration of this analysis, we decided to examine the augmented written genres setup in verbs excluded experiments, with the overall result being 64.18 F1 (to 64.60 F1 non-augmented), which partly contradicts our hypothesis about the specific benefits of the written genres for the nominal coreference resolution task, and probably points to the mere quantitative advantage of Test C for this task. It can again be partially explained by the verbal distribution among the genres, where newswires (nw), which is the dominant basis of the written genre experiments, also are among the most influenced by the verb exclusion procedure.

To see if the verb mention exclusion negatively influences not only the performance on the Twitter set but also on OntoNotess, we excluded verbal mentions from the Official Shared-Task 2012 Test set and predicted it with the model trained on OntoNotes training set annotated only with nominal coreference. The results show a decrease in performance of almost 2% compared to the best results reported in (K. Lee et al. 2018). In comparison to our reproduction experiment (73,84), it loses almost 3%.

The loss is worse in precision than in recall, meaning that the system produces more false-positive annotations, which in turns means that less annotated mentions are correct. Simultaneously, as the recall did not change that much, we can make the hypothesis that in the case of the verb excluded experiment, there was no change in the already correctly annotated mentions (Section 4.3), but that this model added new wrongly annotated ones.

The most influenced metrics among these three are CEAF, an algorithm aligning entities in key and response applying a similarity metric for each pair of entities to measure the correctness of each possible alignment (Cai and Strube 2010). It points to the problem in the chains, or mentions linking in particular. Full results recapitulation can be seen in Table 21.

Table 21. Comparison of the results with/without verbs excluded

	MUC			B3			ceafe			All
Test	Pres.	Recall	F1	Pres.	Recall	F1	Pres.	Recall	F1	F1
Best Lee et al. (2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73
Nominal Onto.	78.84	78.85	78.85	69.68	68.06	68.86	65.2	65.73	65.47	71.06
Diff	**2.56**	0.65	1.55	**2.52**	1.44	1.94	**3**	1.37	**2.13**	1.94

Section 5. Normalization

Another approach to improve the baseline performance is preprocessing or normalization of the data. Text normalization of user-generated content is an important NLP task, especially relevant to the highly sparse and noisy Twitter data. The process itself is used to ‘clean’ the input tokens transforming all non-standard lexical or syntactic variations into their canonical ‘forms’ (Mosquera and Moreda 2013).

5.1 Normalization proceedings

It is important to underline that the data used in the previous experiments was already slightly normalized, through anonymization of parenthesis (‘LBR’ – left bracket – and ‘RBR’ – right bracket), emoji (%EMOJI), and smileys (%SMILEY), which gained 1 percent of performance.
We led several experiments to handle the further noisiness of the data and tried to make it as similar to OntoNotes as possible. The only restriction we imposed on this step was to preserve the number of tokens to prevent severe changes as our data already included gold tokenization and sentence segmentation in CoNLL format.

The normalization was inspired by (Angiani et al. 2016) and (Sidarenka et al. 2013) and overall included:

1. Elimination of all # if they are not tokens by themselves.
2. ‘=‘ equality sign transformation into the verb ‘to be’, third person singular ‘is’.
3. All quotation marks unified as “”.
4. ‘N’t’, is transformed into full negation ‘not’, while ‘’ve’ is transformed into have, ‘ca’ into can, and ‘wo’ into ‘will’ respectively.
5. ‘Vs’ is changed into ‘against’.
6. All repeating interrogations, quotation, and suspension marks are eliminated, leaving one instance.
7. Upper case all ‘i’, which are separate tokens.
8. No repetition of characters more than two times.
9. Words that are fully capitalized are made lowercase, except for organizations and abbreviations like US, UK, WWF.
10. Months got the uppercase first letter.
11. All the links are transformed into ‘.’ punctuation marks.
12. @ user mention mark is treated based on its position and use in the sentence. If it takes part in the syntax of the sentence, the @ mention is eliminated, but the mention itself stays

preserved. If it is at the end and the beginning of the sentence (i.e. it is auto-inserted by Twitter) and does not play any syntactic role, it is treated as links and transformed to '.' .

13. Some of the repeating abbreviation and slang is handled: em, yo/u/ya, yr, cuz transformed into them, you, year, course/cause depending on the context.

14. All smileys and emojis are also treated as links, being transformed into '.' punctuation marks.

5.2 The Normalization results using the Stanford system

The following transformations were applied using e2e-coref and the Stanford neural system, as we used the latter to create the dependency parses and Named Entity tags needed for the experiments described in Section 3 and 4. We also considered it also interesting to compare how two different systems answer to the preprocessing steps, as the core of the Stanford system is parsing information, while e2e-coref uses only token-level flat representations.

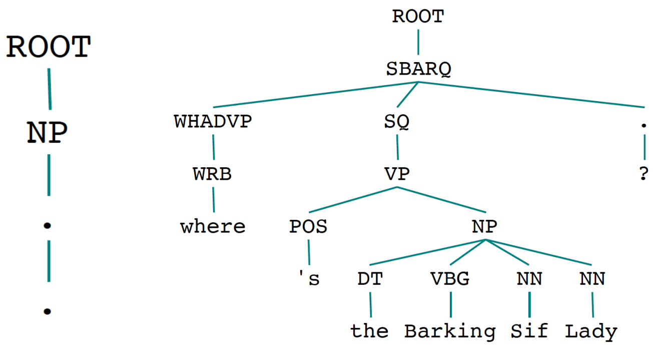
In the context of the Stanford system, there are several options for the prediction file format: txt, XML, and CoNLL. As we found that there is no possibility of introducing speaker information using txt and XML formats, we concentrated on the CoNLL format. At the same time, the tests on CoNLL input where we eliminated all the speaker information and then made all speakers 'PERS0', turned out to be inconclusive, as the difference in the results is around 0.2%, so we reckon that the Stanford system does not benefit from the speaker information. It is also important to mention that in comparison to txt format inputs, the precision drops significantly for the mentions detection from 84.05% to 37.75% when using the CoNLL format, which means that the systems overproduce new wrong mentions. The decline (from 73 to 50.16%) in recall also proves this idea. All this difference is due to the various tokenization and sentences segmentation approaches applied to each format.

There were also several tests concerning parses. It appears that when we impose our gold segmentation, the quality of parses produced by Stanford CoreNLP declines and thus the overall metrics. The best baseline without normalization for the Stanford system was achieved in txt format (43,5% F1), on the 'out-of-the-box' system, without retraining, with the output realigned afterward for the purposes of evaluation. A deeper insight into these results with txt format is discussed in Section 5.4.

5.3 The Normalization results using e2e-coref system

With e2e-coref we were first of all interested in the possibilities of increasing the best result (66.8 F1) of the experiments explained in Sections 3 and 4, namely the ones using training on spoken genres with the aligned Twitter data. We first applied the best preprocessing configuration, based on our experiments with the Stanford system. The following experiments are conducted with normalization of the Twitter data in both training and test sets with the procedure described in 5.1. We had a feeling that converting mentions, links, and other Twitter-specific tokens into periods can be too radical for this system. The Stanford system benefits from it, basically separating this punctuation, especially in the beginning of the sentences into separate parse trees, which only consist of the punctuation and are not considered for the coreference resolution. On Image 1 you can see how a sentence "@mention Where's the Barking Sif Lady" is converted into two different parses and treated separately.

Image 1. Parsed sentence example



The cleaned parses of the sentences become more similar to OntoNotes, which could explain the increase in performance. In the case of e2e-coref, on the contrary, we create sentences starting with punctuational tokens, which is never the case in bc and tc genres.

That is why we decided to repeat this test; however, just anonymizing user mentions into '@mention', also without any conversion of links, emojis, and smileys.

As shown in Table 22, the result of Test 1 is drastically worse than the unprocessed test, with 61.52 versus 66.8 F1 without pre-processing. This first decision to start from the already prepared settings, explain the descending approach chosen for these tests. We do not add step by step new normalization techniques, but on the contrary, take them out one by one, to see which of the steps suitable for the Stanford system, prove to be incompatible with e2e-coref.

In Test 2, we eliminated all the preprocessing of the mentions and hash-tags, and excluded the steps applied to contracted forms of the verbs 'can', 'will not', etc. We slightly gained in performance (61.99 F1) in comparison to Test 1, still being far behind the unprocessed experiment.

In Test 3, we took out lower-casing of all tokens, except for the organization and abbreviations, which again could be a radical step, uniformizing all tokens, and eliminating important discriminative indicators. We also excluded quotation unification at this point, as it was also an ambivalent step for the Stanford system. Thus, Test 3 gained a lot with 64.60 F1, still losing 2.2 percent in comparison to raw data Test. Finally, with Test 4, we excluded repeating letters normalization, only leaving steps 2,4,5,7,10 and 13. The result got a huge decrease to 62.34 F1.

Concluding on these trials, we can say with relative confidence that the normalization can be considered a negative factor for retraining using the same normalization on the in-domain training data and test data. It is probable that as there are several types of data in the training set, the noisiness of the data is a 'flag', which helps the system to recognize the Twitter data, as one which needs a special treatment, and otherwise, gets confused when Twitter data becomes more similar to bc and tcs. In addition, we decided that it would be more compatible with comparing the results of e2e-coref in the same settings, with the model trained exclusively on OntoNotes, without Twitter data in the training set.

Table 22. Normalizing best result experiment

Test	Mention Identification			Chains linking		
Metric	Recall	Precision	F1	Recall	Precision	F1
Test 1	68.13	79.42	73.35	56.8	67.21	61.52
Test 2	71.07	83.09	76.61	57.0	68.0	62.0
Test 3	72.05	84.72	77.88	59.94	70.05	64.6
Test 4	69.6	84.27	76.24	56.99	68.80	62.34

5.4 Comparison

Comparing the two systems' performance on Twitter data, we, first of all, measured its performance without normalization. As we can see in Table 23, while the Stanford system seems to win in terms of mention detection, it loses 2% overall in coreference resolution itself. In general, the Stanford system wins in the recall but loses in precision, which means that it probably tends to overproduce spans, more than e2e-coref. At the same time, for mention identification, the Stanford system has a higher precision, while e2e-coref wins in recall, meaning e2e-coref detects more mentions in general, but that fewer of them are correct.

Table 23. No normalization

Test 0	Stanford neural system			e2e-coref		
Metric	Recall	Precision	F1	Recall	Precision	F1
Mention detection	50,16	**84,05**	**62,83**	**53.43**	71.47	61.15
Chains linking	34,01	**59,22**	43,18	**39.29**	53.89	**45.18**

Reproducing the first test, which included preprocessing steps 1-10, the tendencies found in Test 0 do not change in general, but intensify, with e2e-coref outperforming Stanford system almost by 5% (see Table 24).

Table 24. Normalization. Test 1

Test 1 (1-10)	Stanford neural system			E2e-coref		
Metric	Recall	Precision	F1	Recall	Precision	F1
Mention detection	50,8	**84,16**	**63,35**	**55,88**	70,37	62,29
Chains linking	34,34	**59,27**	43,43	**41,92**	54,05	**46,86**

With Test 2, detailed in Table 25, which included steps 11-12 (links and mentions conversion into punctuation), we see no change in overall tendencies again, with e2e-coref gaining 10% in precision for chain linking, without a change in the recall, and a slight increase in mention identification of 1.5%, due to the similar 10% gain in precision covering up for the slight decrease in the recall, with 50.42% overall (5.24% increase in comparison to raw data test).

At the same time, the Stanford system only gains 2.08% with these normalization steps, with a small overall increase in both recall and precision in both mention identification and chains linking.

Table 25.Normalization. Test 2

Test 2 (1-12)	Stanford neural system			E2e-coref		
Metric	Recall	Precision	F1	Recall	Precision	F1
Mention detection	52,06	85,13	64,61	53.18	79.77	63.82
Chains linking	35,98	61,13	45,26	41,95	64.50	50.42

Finally, the best pre-processing configuration for the Stanford system, which added smileys, emoji and non-standard pronouns normalization, leads to seemingly a statistically irrelevant 0.06% loss, due to the decrease in recall, which is slightly overcoming the gain in precision for both metrics. Interestingly, on the contrary, the Stanford system does not gain much in precision in this setup, but shows an increase in recall, leading to the slight 0.11% gain. Otherwise, we can consider these steps as non-influential. (See Table 26).

Table 26. Normalization. Test 3

Test 2 (1-14)	Stanford neural system			E2e-coref		
Metric	Recall	Precision	F1	Recall	Precision	F1
Mention detection	52,21	85,37	64,8	52.69	80.22	63.6
Chains linking	36.07	61,25	**45,37**	41,66	64,9	50,36

5.5 Normalizing verbal mentions excluded configuration

As the training configuration with verbal mention excluded from OntoNotes has always shown a deviation from the general pattern, we also decided to apply the normalization procedure to this setup.

As we can see from Table 27, similarly to the normalization experiments conducted over the best result (Section 5.3), our normalization steps are not beneficial to the predicting accuracy. The best overall result in terms of both mention detection and coreference resolution is the non-normalized one. However, we can see that the best mention detection precision is obtained with Test 3 with 4.4% gain over the non-normalized setup, which proves that normalization helps to identify the correct mentions better, but at the same time, all the preprocessing tests make the mention identification recall drop, with Test 2 having the most negative impact, which means that normalization leads to excessive loss in a number of the overall found mentions, leaving many behind.

Overall, chain linking also appears to be the best performing in Test 0, although we got an increase in recall with Test 1 (shadowed by the drastic drop in precision) and the best precision with Test 3, again proving itself to be useful to get the correct prediction increase. However, as we can see from Table 27, it continually comes with a recompensation in terms of recall fall.

Table 27. Normalization of the verbal mention excluded configuration. Tests 0-3

Test №	Metric	Recall	Precision	F1
Test 0	Mention detection	**55.63**	82.54	**66.47**
	Chains linking	42.46	64.76	50.99
Test 1	Mention detection	54.65	69.9	61.34
	Chains linking	42.73	56.47	48.47
Test 2	Mention detection	50.24	86.86	63.66
	Chains linking	39.23	70.3	50.01
Test 3	Mention detection	50.73	**86.97**	64.08
	Chains linking	39.58	70.42	50.32

Discussion

This Master thesis covered three approaches of domain adaptation, applying them to increase the performances of a coreference resolution system on Twitter conversations: retraining, annotation schema alignment, and normalization.

In Section 3, we covered the retraining of the system, adding in-domain data to the training set. We also introduced a hand-crafted semi-statistical procedure for the selection of a test set, which is mostly based on linguistic features analysis, as opposed to random sampling, which can be a source of imbalanced evaluation with the relatively small amounts of in-domain data (the case relevant for our study) and domain adaptation tasks in general.

At the same time, instead of the usual choice in modern Machine Learning of training with the maximum possible amount of annotated data, with no regard for its kind and compatibility, we chose to train only on the most statistically and linguistically appropriate sub-corpora of OntoNotes — the reference corpus for coreference resolution in English. This procedure proved beneficial, as the highest performance of our system across all experimental settings was achieved by training only on the spontaneous speech genres, combined with the Twitter corpus training set. This training configuration strongly outperformed all the other setups and brings an improvement of almost 25% over the performance of the system trained only on OntoNotes. More precisely, through a detailed error analysis, we find that overlapping spans in the predicted annotations, which were caused by discrepancies between the annotation schemas used for OntoNotes and the Twitter corpus. It naturally led us to compare the differences in decisions made by the annotators of these two data sources, which we semi-automatically aligned, to exclude this bias from our results (Section 4). Repeating the experiments designed in Section 3 after this alignment step leads to changes in the performance patterns: while the best results are still obtained by training only with the spontaneous speech parts of OntoNotes, augmenting these parts with data from OntoNotes' development and test set brought no improvement. This setup is also the best performing in this study, with 66.8% F1 an increase of 26.9% over the baseline (Section 3). In our interpretation, the experiments with aligned annotations seem to rely more on the in-domain data. When we applied the default model, trained on OntoNotes, the alignment of the schemas gained a vast improvement of 6%. An important alignment step left apart in the setup above concerns verbal mentions, which are considered included in OntoNotes but not in the Twitter corpus, which only focuses on the nominal coreference resolution task. As there is no way to automatically annotate verbs in the Twitter corpus, we automatically excluded them in the OntoNotes instead. Thus, our decision to make it a separate experiment is motivated by this intrusion into OntoNotes, making it less comparable to the previous experiments.

Interestingly, by excluding the verb mentions, we gained 10% for the model trained on the full OntoNotes corpus. However, we report considerable losses for all the configurations, which include Twitter data in the training set, and especially with the spoken genres, which, as our quantitative analysis suggests, lose the biggest amount of annotations after this procedure due to their conversational nature.

Hence, after all, we can conclude that the annotation of verbal mentions in OntoNotes is beneficial for our task.

Finally, in Section 5, we concentrated on the Normalization, tested before with the Stanford neural system, which is also trained on the OntoNotes, to try to augment further the best result we have. However, we found out that normalization does not have any positive influence when applied to the Twitter data in both training and test set, and, on the contrary, leads to the considerable performance decrease. We reckon that the noisiness of the Twitter data in these configurations can be used by the model to identify tweets better to treat them slightly differently from the spoken genres, and consequently, losing these 'flags' with normalization explains the performance losses.

Thus, we decided to compare the Stanford system and e2e-coref package models, both only trained with OntoNotes, reaction to the same normalization steps. We report that the e2e-coref 'out-of-the-box' model not only outperforms the Stanford neural system with and without normalization but also gains more from our pre-processing procedure, increasing performance by 5% (with only around 3% in case of Stanford neural system). Conversion of the auto-inserted user mentions into the period, and non-standard pronouns correction proved to be the most beneficial for both system's prediction accuracy. The verbal mention excluded experiment trained on full OntoNotes shows a change in the pattern again, with normalization having no positive influence on the overall results, with Test 3 including the full list of the pre-processing steps, showing precision improvement for both mention identification and chain linking, compensated, however, by the drop of the recall.

Conclusion

Twitter conversations proved to be a complicated material for coreference resolution processing. By adding in-domain training data, and selecting the most suitable OntoNotes genres, we have improved the performance of a state-of-the-art coreference resolution system by 26%, with our best result still being **66.8%**, which is 13% lower of the current state-of-the-art system and 6% lower than the performance ours, both evaluated on the OntoNotes official evaluation test set. This proves that even a small amount of in-domain data in the training set can have a meaningful impact, and more generally, that quantity over quality is not always the best decision for Deep Learning systems. Through error analysis, we also showed that this improvement is not the result of a large addition of the first and second person pronouns but of a qualitative improvement in the resolution of third person pronouns.

We have conducted several trials, excluding verbal mentions from the OntoNotes,, with our results showing different patterns from the previous results, suggesting that further studies in this direction are needed. Finally, applying normalization on the Twitter dataÂ led us to a further 5% increase in performance, though only while training on the full OntoNotes training set, without adding in-domain data, while our best setup, which includes spoken genres and Twitter corpus failed to show any improvement in this setting.

We consider that this study is of both relevance and significance, as it is one of the first of its kind in terms of domain adaptation of coreference resolution systems to Twitter conversations. This genre type isÂ especially important nowadays, with public sentiment and opinion analysis studies relying more and more on data mined from tweets. Coreference resolution may be the way of providing downstream systems with a deeper comprehension of the complicated relations between the referenced concepts and their evolution in the conversation. In the end, it can lead us to a broader linguistic understanding of modern languages and even social psychology, with Twitter being its infinite source today.

Limitations and future work

There are evidently many more possibilities of normalization left to explore. Follow-up work could concentrate upon learning normalization methods (Muller, Sagot, and Seddah 2019), which would, however, presuppose a retokenization and will be optimally done directly on the original MMAX2 (annotation) files used, with the pipeline to recreate golden segmentation and tokenization established. The normalization steps we used were all constrained by the idea of not changing the number of tokens, and thus were not necessarily optimal for this task.

The most confusing experimental setup among those we present is the one following the verb mentions exclusion, which suggests that the pattern of performance loss associated with the number of the verb mentions excluded definitely deserves a closer look.

In the context of the retraining, some additional hyper-parameters tuning could be applied to the best result we reached, in order to optimize the system to the new type of data. However, it would require the creation of a development set, in order to overfit the test set, which is already slightly biased due to our choice of handcrafted — rather than random — sampling.

From the linguistic perspective, more experiments with training data combinations could provide deeper insights. In particular, beyond the spoken/written paradigm, the divide between formal and informal language seems promising.

In the Related work section, we have already discussed some soft constraint algorithms based on domain features applied by other authors, which could help to get better results without retraining the system, which, combined with in-domain data, may lead to even better results, optimizing out-of-domain data of the training set.

Having access to more computational power would also allow us to test more recent state-of-the-art systems, and in particular those relying on contemporary contextual embeddings.

Acknowledgments

To Manfred Stede, Berfin Aktaş, and Annalena Kohnert for their constant support, as well as their intellectual and creative input into the design of the experiments and their analysis.

As part of this Master thesis is subject to the collaboration on the paper ‘Adapting Coreference Resolution to Twitter Conversations ’submitted to EMNLP 2020 and currently being peer-reviewed and still unpublished, this Master thesis can be considered its extension, including also the preparatory and follow-up work of the one described in the paper, but it cannot directly cite the paper as it precedes the paper’s hopefully eventual appearance.

To Loïc Grobol for his infinite patience and optimism during proof-reading of this thesis.

Bibliography

Aktaş, Berfin, Tatjana Scheffler, and Manfred Stede. 2018. "Anaphora Resolution for Twitter Conversations: An Exploratory Study." *Computational Models of Reference, Anaphora and Coreference*. https://developer.twitter.com/en/docs/tweets/data-.

Aktaş, Berfin, Tatjana Scheffler, and Manfred Stede. 2019. "Coreference in English OntoNotes: Properties and Genre Differences." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-27947-9_15.

Angiani, Giulio, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi. 2016. "A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter." In *CEUR Workshop Proceedings*.

Apache Software Foundation. 2014. "Apache OpenNLP Natural Language Processing Library." 2014. http://opennlp.apache.org/.

Bagga, Amit, and Breck Baldwin. 1998. "Algorithms for Scoring Coreference Chain." *First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*.

Cai, Jie, and Michael Strube. 2010. "Evaluation Metrics for End-to-End Coreference Resolution Systems." In *Proceedings of the SIGDIAL 2010 Conference: 11th Annual Meeting of the Special Interest Group On Discourse and Dialogue*.

Clark, Kevin, and Christopher D. Manning. 2015. "Entity-Centric Coreference Resolution with Model Stacking." In *ACL-IJCNLP 2015 - 5third Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/v1/p15-1136.

Clark, Kevin, and Christopher D. Manning. 2016a. "Deep Reinforcement Learning for Mention-Ranking Coreference Models." In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/d16-1245.

Clark, Kevin, and Christopher D. Manning. 2016b. "Improving Coreference Resolution by Learning Entity-Level Distributed Representations." In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*. https://doi.org/10.18653/v1/p16-1061.

Comrie, Bernard, R. E. Asher, and J. M. Y. Simpson. 1995. "The Encyclopedia of Language and Linguistics." *Language*. https://doi.org/10.2307/415969.

Connolly, Dennis, John D. Burger, and David S. Day. 1994. "A Machine Learning Approach to Anaphoric Reference."

Deemter, Kees Van, and Rodger Kibble. 2000. "On Coreferring: Coreference in MUC and Related Annotation Schemes." *Computational Linguistics*. https://doi.org/10.1162/089120100750105966.

Durrett, Greg, David Hall, and Dan Klein. 2013. “Decentralized Entity-Level Modeling for Coreference Resolution.” In *ACL 2013 - 5first Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

Durrett, Greg, and Dan Klein. 2013. “Easy Victories and Uphill Battles in Coreference Resolution.” In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Grosz, Barbara J. 1977. “The Representation and Use of Focus in Dialogue Understanding, Technical Report No.151.” Menlo Park, California: SRI International.

Hajič, Jan, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. 2017. “Prague Dependency Treebank.” In *Handbook of Linguistic Annotation*. https://doi.org/10.1007/978-94-024-0881-2_21.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation*. https://doi.org/10.1162/neco.1997.9.8.1735.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. “OntoNotes: The 90% Solution.” In *Human Language Technology Conference of the NAACL, Short Papers*.

Iida, Ryu, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. “Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations.” In *ACL 2007: The LAW - Proceedings of The Linguistic Annotation Workshop*. https://doi.org/10.3115/1642059.1642081.

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. “SpanBERT: Improving Pre-Training by Representing and Predicting Spans.” *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00300.

Joshi, Mandar, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2020. “BERT for Coreference Resolution: Baselines and Analysis.” In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/d19-1588.

Kantor, Ben, and Amir Globerson. 2020. “Coreference Resolution with Entity Equalization.” In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. https://doi.org/10.18653/v1/p19-1066.

Karttunen, Lauri. 2020. “Discourse Referents.” In *Notes from the Linguistic Underground*. https://doi.org/10.1163/9789004368859_021.

Kenton, Ming-wei Chang, Lee Kristina, and Jacob Devlin. 2017. “BERT Paper.” *ArXiv:1810.04805 [Cs]*. https://doi.org/arXiv:1811.03600v2.

Kingma, Diederik P., and Jimmy Lei Ba. 2015. “Adam: A Method for Stochastic Optimization.” In *third International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kübler, Sandra, and Desislava Zhekova. 2011. “Singletons and Coreference Resolution Evaluation.” In *International Conference Recent Advances in Natural Language Processing, RANLP*.

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. “Stanford ’ s Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task.” In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics*.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. “End-to-End Neural Coreference Resolution.” In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/d17-1018.

Lee, Kenton, Luheng He, Luke Zettlemoyer, and Paul G Allen. 2018. “Higher-Order Coreference Resolution with Coarse-to-Fine Inference.”

Luo, Xiaoqiang, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. “A Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree.” In . https://doi.org/10.3115/1218955.1218973.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David Mcclosky. n.d. “The Stanford CoreNLP Natural Language Processing Toolkit.”

Mosquera, Alejandro, and Paloma Moreda. 2013. “Improving Web 2.0 Opinion Mining Systems Using Text Normalisation Techniques.” In *International Conference Recent Advances in Natural Language Processing, RANLP*.

Muller, Benjamin, Benoit Sagot, and Djamé Seddah. 2019. “Enhancing BERT for Lexical Normalization.” In . https://doi.org/10.18653/v1/d19-5539.

Nair, Vinod, and Geoffrey E. Hinton. 2010. “Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair.” In *Proceedings of ICML*.

Ng, Vincent. 2010. “Supervised Noun Phrase Coreference Research: The First Fifteen Years.” Association for Computational Linguistics. http://www.itl.nist.gov/iad/mig/tests/ace/.

Ngoc, Quynh, Thi Do, Steven Bethard, and Marie-Francine Moens. 2015. “Adapting Coreference Resolution for Narrative Processing.” Association for Computational Linguistics. http://dspace.mit.edu/handle/1721.1/57507.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/v1/d14-1162.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. “Deep Contextualized Word Representations.” In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. https://doi.org/10.18653/v1/n18-1202.

Poesio, Massimo, and Ron Artstein. 2008. "Anaphoric Annotation in the ARRAU Corpus." In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. n.d. "Towards Robust Linguistic Analysis Using OntoNotes."

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes." http://www.bbn.com/nlp/ontonotes.

Pradhan, Sameer, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. "Unrestricted Coreference: Identifying Entities and Events in OntoNotes." *ICSC 2007 International Conference on Semantic Computing*, 446–53. https://doi.org/10.1109/ICSC.2007.93.

Press, C U. 2008. *Cambridge Advanced Learner's Dictionary*. Cambridge University Press. https://books.google.com.ua/books?id=Rqg0RTroDK8C.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2015. Adapting coreference resolution for narrative processing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2262–2267, Lisbon, Portugal. Association for Computational Linguistics.

Raghunathan, Karthik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. "A Multi-Pass Sieve for Coreference Resolution." In *EMNLP 2010 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Recasens, Marta, and M. Antònia Martí. 2010. "AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan." *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-009-9108-x.

Ritter, Alan, Sam Clark, and Oren Etzioni. 2011. "Named Entity Recognition in Tweets: An Experimental Study." Association for Computational Linguistics.

Sidarenka, Uladzimir, Scheffler, Tatjana and Stede, Manfred. Rule-based Normalization of German Twitter Messages. In Proceedings of the Conference of the German Society for Computational Linguistics (GSCL 2013). Darmstadt, Germany, September 2013. European Language Resources Association (ELRA).

Sidner, Candace Lee. 1979. "Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse." *Massachusetts Inst of Tech Cambridge Artificial Intelligence Lab*. https://doi.org/10.1007/BF01250453.

Stede, Manfred. 2011. "Discourse Processing." *Synthesis Lectures on Human Language Technologies* 4 (3): 1–165. https://doi.org/10.2200/S00354ED1V01Y201111HLT015.

Sukthanker, Rhea, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2018. “Anaphora and Coreference Resolution: A Review.”

Telljohann, Heike, Erhard Hinrichs, and Sandra Kübler. 2004. “The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone.” In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. 2010. “Word Representations: A Simple and General Method for Semi-Supervised Learning.” In *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

Uryupina, Olga, and Massimo Poesio. n.d. “Domain-Specific vs. Uniform Modeling for Coreference Resolution.” http://sourceforge.net/projects/carafe.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. “A Model-Theoretic Coreference Scoring Scheme.” In MUC6 '95: Proceedings of the 6th conference on Message understanding, pages 45–52. https://doi.org/10.3115/1072399.1072405.

Weischedel, Ralph, Sameer Pradhan, Lance Ramshaw, Michelle Franchini, Mohammed El-bachouti, Martha Palmer, et al. 2010. “OntoNotes Release 4.0.” *Sinorama*.

Weischedel, Ralph, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. “OntoNotes: A Large Training Corpus for Enhanced Processing.” *Handbook of Natural Language Processing and Machine Translation. Springer*.

Wiseman, Sam, Alexander M. Rush, Stuart M. Shieber, and Jason Weston. 2015. “Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution.” In *ACL-IJCNLP 2015 - 5third Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/v1/p15-1137.

Zitouni, Imed, Xiaoqiang Luo, and Radu Florian. 2010. *Arabic Computational Linguistics. A Statistical Model for Arabic Mention Detection and Chaining. In Farghaly, A.* CSLI Publications, Center for the Study of Language and Information. https://web.stanford.edu/group/cslipublications/cslipublications/site/9781575865447.shtml.

Appendixes

Appendix 1. Baseline experiment statistics

Genre	Tokens	Chains	Mentions
Test A (OntoNotes only)			
Bc	144K	4236	18K
Bn	172K	6138	21K
Mz	164K	3534	13K
Nw	387K	9404	34K
Pt	210K	6611	42K
Tc	81K	1931	12K
Wb	131K	2993	12K
Total	1289K	34K	152K
per cent of internet genres (twitter/wb)	10,16 %	8,8 %	7.9%
percent of spoken genres (bc, tc)	18 %	19 %	17,15 %
Result in F1	39,77 %		
Test B (OntoNotes, Twitter)			

Twitter	44912	6589	1596
OntoNotes	1289K	34K	152K
Total	1268785	202970	87452
per cent of twitter/ total	3,55 %	4 %	2,5 %
per cent of internet genres (twitter/wb)	10,3 %	11,8 %	10,95 %
per cent of spoken genres (bc,tc)	17,6 %	18 %	16,7 %
Result in F1	59,5633 %		
Test C (bc, tc, wb, Twitter)			
Twitter	44912	6589	1596
Bc	144K	4236	18K
Tc	81K	1931	12K
Wb	131K	2993	12K
Total	400912	15749	43596
per cent of twitter/ total	11 %	13 %	9 %
per cent of internet genres (twitter/wb)	36 %	39 %	39,5 %
per cent of spoken genres (bc, tc)	64 %	61 %	60,5 %

Result in F1	60,37 %		
Test D (bc, tc, Twitter)			
Twitter	44912	6589	1596
Bc	144K	4236	18K
Tc	81K	1931	12K
Total	269912	12756	31596
per cent of twitter/ total	16,9 %	17,6 %	13 %
per cent of spoken genres (bc, tc)	83,1 %	82,4 %	87 %
Result in F1	64,27 %		
Test E (mz, part of nw, twitter)			
Twitter	44912	6589	1596
Mz	164K	3534	13K
Nw	60K	5500	16K
Total	269K	15623	30K
per cent of twitter/ total	16,8 %	19 %	12,2 %
per cent of written genres (bc, tc)	83,2 %	81 %	87,8 %

Result in F1	62,39 %		
Test F (Twitter only)			
Twitter	44912	6589	1596
Total	Twitter	44912	6589
per cent of twitter/ total	100	100	100
Result in F1	61,86 %		
Test G (Twitter; bc, tc, augmented by 40 %)			
Twitter	44912	6589	1596
Bc+tc	315 615	50155	19343
Total	360527	58086	21510
per cent of twitter/ total	12,45 %	13,65 %	10 %
per cent of spoken genres (bc,tc)	87,55 %	86,35 %	90 %
Result in F1	65,6 %		
Test H (Twitter; mz, nw, wb, augmented by 40 %)			
Twitter	44912	7931	2167
Mz	164096	18438	8731

Nw+wb	151584	24322	11204
Total	360592	50691	22102
Percent of twitter/ total	12,45	15,6	9,8
percent of written genres (bc,tc)	87,55	84,4	90,2
Result in F1	61,41 %		

Appendix 2. Baseline experiment error analysis: Mentions

	Test A	Test G
All predicted mentions		
Number mentions	305	377
Mean mention length (in tokens)	1.6032786885245902	1.4721485411140585
Number Pronouns	168	227
first Person Pronouns	40	60
second Person Pronouns	29	72
third Person Pronouns	74	69
Number non-standard Pronouns	5	7
@-mentions	51	3
Tweet-initial @-mentions	46	0
Hashtag-mentions	0	6
Mentions including hash-tag	0	7
Found gold mentions		
Number mentions	212	334

Mean mention length (in tokens)	1.25	1.377245508982036
Number Pronouns	150	215
first Person Pronouns	38	56
second Person Pronouns	26	68
third Person Pronouns	62	65
Number non-standard Pronouns	5	7
@-mentions	5	3
Tweet-initial @-mentions	1	0
Hashtag-mentions	0	5
Mentions including hash-tag	0	6
Missed gold mentions		
Number mentions	250	128
Mean mention length (in tokens)	2.032	2.4453125
Number Pronouns	82	18
first Person Pronouns	21	3
second Person Pronouns	42	1

third Person Pronouns	8	5
Number non-standard Pronouns	2	0
@-mentions	3	5
Tweet-initial @-mentions	0	1
Hashtag-mentions	11	6
Mentions including hash-tag	13	7
Extra predicted mentions		
Number mentions	93	43
Mean mention length (in tokens)	2.4086021505376345	2.2093023255813953
Number Pronouns	17	12
first Person Pronouns	2	4
second Person Pronouns	3	4
third Person Pronouns	12	4
Number non-standard Pronouns	0	0
@-mentions	46	0
Tweet-initial @-mentions	45	0

Hashtag-mentions	0	1
Mentions including hash-tag	0	1

Appendix 3. Baseline experiment error analysis: Chains

	Test A	Test G
All Predicted Chains		
Number Chains	110	111
Mean chain length	2.772727272727273	3.3963963963963963
Mean mention length (in tokens)	1.6956709956709957	1.6062224562224563
Number pronoun chains	23	27
Non-standard pronouns	5	7
Found gold chains		
Number Chains	14	41
Mean chain length	2.5	2.6341463414634148
Mean mention length (in tokens)	1.5952380952380951	1.5243321718931477
Number pronoun chains	4	13
Non-standard pronouns	0	0
Missed entities		
Number Chains	51	22

Mean chain length	2.4705882352941178	2.272727272727273
Mean mention length (in tokens)	2.0954715219421103	2.5757575757575757
Number pronoun chains	4	1
Non-standard pronouns	1	0
Completely predicted Chains with additional mentions		
Number Chains	9	18
Mean chain length	4.444444444444445	9.38888888888889
Mean mention length (in tokens)	1.3539682539682538	1.3509499759499761
Number pronouns in additional mentions	10 / 17 0.5882352941176471 %	79 / 105 0.7523809523809524 %
Additional mentions in chain are not in other chain	1	4
All additional mentions are pronouns	4	6
Additional mentions in chain are coreferent	8	5
Additional mentions in chain are pronouns and in no gold chain	0	3
Additional mentions in chain are pronouns and coreferent	4	2
Chains with firstPersonSG mismatch(es)	0	7