Enriching Basque Coreference Resolution System using Semantic Knowledge sources

Ander Soraluze and **Olatz Arregi** and **Xabier Arregi** and **Arantza Díaz de Ilarraza**
University of the Basque Country
Donostia - San Sebastián, Spain
{ander.soraluze,olatz.arregi, xabier.arregi,a.diazdeilarraza}@ehu.eus

Abstract

In this paper we present a Basque coreference resolution system enriched with semantic knowledge. An error analysis carried out revealed the deficiencies that the system had in resolving coreference cases in which semantic or world knowledge is needed. We attempt to improve the deficiencies using two semantic knowledge sources, specifically Wikipedia and Word-Net.

1 Introduction

Coreference resolution consists of identifying textual expressions (mentions) that refer to real-world objects (entities) and determining which of these mentions refer to the same entity. While different string-matching techniques are useful to determine which of these mentions refer to the same entity, there are cases in which more knowledge is needed, that is the case of the Example in 1.

(1) [**Osasunak**] lehenengo mailara igotzeko lehian azken astean bizi duen giroa oso polita da. [**Taldea**] lasaitzeko asmoz Oronozera eraman zituen Lotinak atzo guztiak. Oronozko kontzentrazioa beharrezkoa dute [**gorritxoek**].

*"[**Osasuna**] is going through a beautiful moment in the last week in the race to ascend to the Premier League. In order to reassure [**the team**] Lotina has decided to give all of them to Oronoz. [**The reds**] need to concentrate in Oronoz."*

Having the world knowledge that *Osasuna* is a *football team* and its nickname is *the reds* would be helpful for establishing the coreference relations between the mentions [**Osasuna**], [**Taldea**] and [**gorritxoek**] in the example presented above.

Evaluation scores used in coreference resolution tasks can show how effective a system is; however, they neither identify deficiencies of the system, nor give any indication of how those errors might be corrected. Error analyses are a good option that can help to clear the deficiencies of a coreference resolver. Bearing this in mind, we have carried out an error analysis of the extended version of the coreference resolution system presented in Soraluze et al. (2015). In this paper we present an improvement of this Basque coreference resolution system by using semantic knowledge sources in order to correctly resolve cases like in Example 1.

This paper is structured as follows. After presenting an error analysis of the coreference resolution system in Section 2, we analyse similar works to ours in which semantic knowledge sources have been used to improve coreference resolution in Section 3. Section 4 presents how we integrated the semantic knowledge in our system. The main experimental results are outlined in Section 5 and discussed in Section 6. Finally, we review the main conclusions and preview future work.

2 Error Analysis

A deep error-analysis can reveal the weak points of the coreference resolution system and help to decide future directions in the improvement of the system. The system we have evaluated is an adaptation of the Stanford Coreference resolution system (Lee et al., 2013) to the Basque language. The Stanford coreference resolution module is a deterministic rule-based system which is based on ten independent coreference models or sieves that are precision-oriented, i.e., they are applied sequentially from highest to lowest precision. All the sieves of the system have been modified taking into account the characteristics of the Basque lan-

guage and, one new sieve has been added, obtaining an end-to-end coreference resolution system.

The corpus used to carry out the error analysis is a part of EPEC (the Reference Corpus for the Processing of Basque) (Aduriz et al., 2006). EPEC is a 300,000 word sample collection of news published in *Euskaldunon Egunkaria*, a Basque language newspaper. The part of the corpus we have used has about 45,000 words and it has been manually tagged at coreference level by two linguists (Ceberio et al., 2016). First of all, automatically tagged mentions obtained by a mention detector (Soraluze et al., 2016) have been corrected; then, coreferent mentions have been linked in clusters.

More detailed information about the EPEC corpus can be found in Table 1.

	Words	Mentions	Clusters	Singletons
Devel	30434	8432	1313	4383
Test	15949	4360	621	2445

Table 1: EPEC corpus division information

2.1 Error types

The errors have been classified following the categorization presented in Kummerfeld and Klein (2013). The tool[1] presented in the paper has been used to help in identifying and quantifying the errors produced by the coreference resolution system:

- **Span Error (SE):** A mention span has been identified incorrectly.

- **Conflated Entities (CE):** Two entities have been unified creating a new incorrect one.

- **Extra Mention (EM):** An entity includes an incorrectly identified mention.

- **Extra Entity (EE):** An entity which consists of incorrectly identified mentions is outputted by the system.

- **Divided Entity (DE):** An entity has been divided in two entities.

- **Missing Mention (MM):** A not identified mention is missing in an entity.

- **Missing Entity (ME):** The system misses an entity which is present in the gold standard.

The error types are summarised in Table 2.

---

[1]code.google.com/p/berkeley-coreference-analyser/

Error Type	System	Gold
Span Error	s_1	s_1 s_2
Conflated Entities	$\{m_1, m_2\}_{e1}$	$\{m_1\}_{e1}$
	-	$\{m_2\}_{e2}$
Extra Mention	$\{m_1, m_2\}$	$\{m_1\}$
Extra Entity	$\{m_1, m_2\}$	-
Divided Entity	$\{m_1\}_{e1}$	$\{m_1, m_2\}_{e1}$
	$\{m_2\}_{e2}$	-
Missing Mention	$\{m_1\}$	$\{m_1, m_2\}$
Missing Entity	-	$\{m_1, m_2\}$

Table 2: Error types. s=string, m=mention, e=entity

2.2 Error causes

Apart from classifying the errors committed by the coreference resolution system, it is important to observe the causes of these error types. These are the causes of errors we found:

- **Preprocessing (PP):** Errors in the preprocessing step (lemmatization, PoS tagging, etc.) provoke incorrect or missing links in coreference resolution.

- **Mention Detection (MD):** These errors are provoked due to incorrectly identified (not a mention, incorrect boundaries..) or missed mentions during mention detection step. Missed mentions directly affect the recall of the system, and incorrectly identified mentions affect precision.

- **Pronominal Resolution (PR):** The system often generates incorrect links between the pronoun and its antecedent.

- **Ellipsis Resolution (ER):** Elliptical mentions do not provide much information as they omit the noun, as a consequence it is difficult to correctly link these types of mentions with their correct antecedent.

  For example, it is complicated to link the elliptical mention [Yosi Beilin Israelgo Justizia ministroak Jeruralemi buruz esandako-Ø[2]-ak] "what Yosi Beilin Israel Justice Minister said" with its antecedent [Beilin Justizia ministroaren hitzak] "Beilin Justice minister's words".

- **Semantic Knowledge (SK):** Errors related to a semantic relation (synonymy, hyperonymy, metonymy) between the heads of two mentions.

---

[2]In this case Ø refers to "what".

For example, in mentions [Libanoko Parlamentuak] "Lebanon parliament" and [Libanoko Legebiltzarrak] "Lebanon parliament", *parlamentua* is a synomyn of *legebiltzarra*.

- **World Knowledge (WK):** In some cases the system is not able to link mentions as a consequence of the lack of world knowledge required to resolve them correctly.

  For example, to link the mention [Reala] "Reala" with the mention [talde txuri-urdinak] "white-blue team", it is necessary to know that *Reala* is a team and the nickname of the football team is *txuri-urdinak* "white-blue".

- **Miscellaneous (MISC):** In this category we classify the errors that are not contained in the above categories.

  An example of a miscellaneous error could be the following. The mention [Kelme, Euskaltel eta Lampre] should be linked with the mention [Hiru taldeak] "The three teams". In this specific example it is necessary to know that Kelme, Euskaltel and Lampre are teams and the enumerated mention has three elements.

After defining the error types and the error causes, we analysed how the error causes affect the error types in EPEC corpus. The distribution of errors is shown in Figure 1.

As we observe in Figure 1, the most common errors types of the system fail in Span Error (29.36%), Conflated Entities (11.92%), Divided Entities (42.88%) and Missing Mention (11.92%) categories.

Observing the error causes, we can conclude that mention detection is crucial for coreference resolution, 52.52% of errors. Improving mention detection would likely improve the scores obtained in coreference resolution. Nevertheless, in order to identify deficiencies of a coreference resolution system, Pronominal Resolution (9.17%), Ellipsis Resolution (3.21%), Semantics (6.42%) and World Knowledge (9.86%) categories can reveal how the errors might be corrected. Due to the variety of errors classified in miscellaneous category, little improvement would be achieved despite making a big effort to solve them.

Among all the error causes, in this paper we are going to focus on errors provoked by the lack of semantic and world knowledge.

3 Related Work

Lexical and encyclopedic information sources, such as WordNet, Wikipedia, Yago or DBPedia have been widely used to improve coreference resolution.

WordNet (Fellbaum, 1998) is the one of oldest resources for lexical knowledge. It consists of *synsets*, which link synonymous word senses together. Using WordNet's structure, it is possible to find synonyms and hyperonymic relations. Wikipedia is a collaborative open source encyclopedia edited by volunteers and provides a very large domain-independent encyclopedic repository. Yago (Suchanek et al., 2007) is a knowledge base, linking Wikipedia entries to the WordNet ontology. And finally, DBPedia (Mendes et al., 2012) contains useful ontological information extracted from the data in Wikipedia.

Regarding works in which lexical and encyclopedic information sources have been exploited, Ponzetto and Strube (2006) were the earliest to use WordNet and Wikipedia.

Uryupina et al. (2011) extracted semantic compatibility and aliasing information from Wikipedia and Yago and incorporated it in coreference resolution system. They showed that using such knowledge with no disambiguation and filtering does not bring any improvement over the baseline, whereas a few very simple disambiguation and filtering techniques lead to better results. In the end, they improve their system's performance by 2-3 percentage points.

Rahman and Ng (2011) used Yago to inject knowledge attributes in mentions, but noticed that knowledge injection could be noisy.

Durrett and Klein (2013) observed that the semantic information contained even in a coreference corpus of thousands of documents is insufficient to generalize to unseen data, so system designers have turned to external resources. Using specialised features, as well as WordNet-based hypernymy and synonymy and other resources, they obtained a gain from 60.06 in CoNLL score to 61.58 using automatic mentions, and from 75.08 to 76.68 with gold mentions.

Ratinov and Roth (2012) extract attributes from Wikipedia pages which they used to improve the recall in their system, based on a hybrid (Lee et al., 2013).
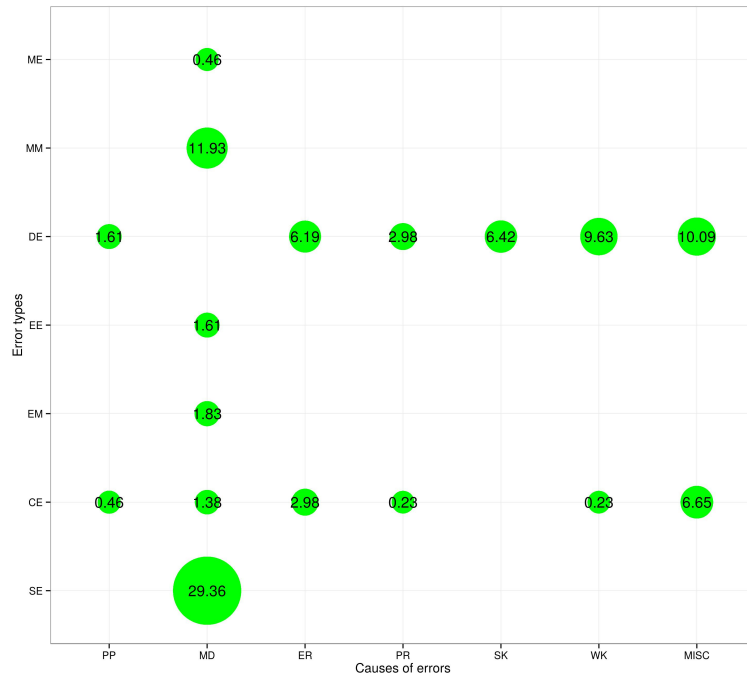
Figure 1: Distribution of error causes into error types.

In Hajishirzi et al. (2013) NECo, a new model for named entity linking and coreference resolution, which solves both problems jointly, reducing the errors of each is introduced. NECo extends the Stanford deterministic coreference resolution system by automatically linking mentions to Wikipedia and introducing new sieves which profit from information obtained by named entity linking.

As pointed out in Recasens et al. (2013), opaque mentions (mentions with very different words like *Google* and *the search giant*) account for 65% of the errors made by state-of-the-art systems, so to improve coreference scores beyond 60-70% it is necessary to make better use of semantic and world knowledge to deal with non-identical-string coreference. They use a corpus of comparable documents to extract aliases and they report that their method not only finds synonymy and instance relations, but also metonymic cases. They obtain a gain of 0.7% F1 score for the CoNLL metric using gold mentions.

Lee et al. (2013) mention that the biggest challenge in coreference resolution, accounting for 42% of errors in the state-of-the art Stanford system, is the inability to reason effectively about background semantic knowledge.

The intuition behind the work presented in Dur-

rett and Klein (2014) is that named entity recognition on ambiguous instances can obtain benefit using coreference resolution, and similarly can benefit from Wikipedia knowledge. At the same time, coreference can profit from better named entity information.

4 Improving Coreference Resolution with Semantic Knowledge sources

This section explains the improvement process of the coreference resolution system with semantic knowledge sources. In order to treat cases where knowledge is needed, two new specialised sieves have been added to the coreference resolution system: One to extract knowledge from Wikipedia and the other to obtain semantic information from WordNet.

4.1 Enriching mentions with Named Entity Linking

Named Entity Linking is the task of matching mentions to corresponding entities in a knowledge base, such as Wikipedia.

As pointed out in Versley et al. (2016), named entity linking, or disambiguation of entity mentions, is beneficial to make full use of the information in Wikipedia.

The Basque version of Wikipedia, contained

11

about 258,000 articles in September 2016, which is much smaller in size when compared with English Wikipedia, which contained about 5,250,837 pages on the same date.

In order to disambiguate and link mentions to Basque Wikipedia pages, the following formula has been applied to all the named entity mentions in a document:

$$P(s, c, e) = P(e \mid s)P(e \mid c)$$

$P(e \mid s)$ is the probability of being entity e given s string, i.e., the normalised probability of being entity e linked with string s in Wikipedia. $P(e \mid c)$ is the probability of being entity e given the context c. The context c is a window of size $[-50, +50]$ of the string s. To calculate $P(e \mid c)$ probability, UKB[3] software has been utilised. UKB software uses *Personalized Page Rank* algorithm presented in (Agirre and Soroa, 2009) and (Agirre et al., 2014) to estimate the probabilities.

If a named-entity mention is linked with any page from Wikipedia, the page that UKB says it is the most probable is used to enrich the mention. From the Wikipedia page the following information is obtained:

- The title of the page. The title sometimes gives useful information. For example, for the named-entity mention *AEK*, the title of its Wikipedia page is *Alfabetatze Euskalduntze Koordinakundea* "Literacy and Euskaldunization Coordinator", where the extent of the acronym is obtained. Furthermore it gives the information that AEK is a coordinator, *koordinakundea*.

- The first sentence. The first paragraph of each Wikipedia article provides a very brief summary of the entity. Usually the most useful information is in the first sentence, this is where the entity is defined.

- If the Wikipedia page has an Infobox, we extract information from it. Infoboxes contain structured information in which the attributes of many entities are listed in a standardized way.

After the information is obtained from the Wikipedia page, this information is processed and the NPs are extracted.

These NPs and their sub-phrases are used to enrich the mentions with world knowledge. To further reduce the noise, the NPs that are location named-entities in a Wikipedia page about a location are discarded.

Taking Example 1, the mention *Osasuna* is enriched as follows: The most probable Wikipedia page proposed by UKB for the mention *Osasuna* is *Osasuna_futbol_kluba* "Osasuna football club". Therefore, we obtain from this page the title, the first sentence and Infobox information. The NPs obtained after the information is processed are *gorritxoak* "the reds", *Osasuna futbol kluba* "Osasuna football club" and *Nafarroako futbol taldea* "football team from Navarre". So the mention *Osasuna* is enriched with the set of lemmas of the NPs and the lemmas of their sub-phrases: {gorritxo, Osasuna futbol klub, futbol klub, klub, Nafarroa futbol talde, futbol talde, talde} "{the reds, Osasuna football club, football club, club, football team from Navarre, football team, team}".

4.2 Wiki-alias sieve

The new Wiki-alias sieve uses the mentions enriched by information obtained from Wikipedia pages.

Using this information, the Wiki-alias sieve assumes that two mentions are coreferent if one of the two following conditions is fulfilled:

i) the set of enriched word lemmas in the potential antecedent has all the mention candidate's span lemmas. To better understand this constraint, suppose that the mention *Realak* is enriched with {talde, futbol talde, txuri-urdin} "{team, football team, white and blue}", as the potential antecedent *Realak* has all the lemmas in the mention candidate's span, i.e., *talde* "{team}" and *txuri-urdin* "{white and blue}", the mention *talde txuri-urdinak* "{white and blue team}" is considered coreferent of *Realak*.

ii) the head word lemma of the mention candidate is equal to the head word lemma of the potential antecedent or equal to any lemma in the set of enriched lemmas of the potential antecedent, and all the enriched lemmas of the potential antecedent appear in the cluster lemmas of the mention candidate. For example, this constraint considers coreferent the potential antecedent *Jacques Chiracek* and the mention candidate *Jacques Chirac Frantziako errepublikako*

presidentea. After *Jacques Chiracek* mention has been enriched with lemmas {presidente, Frantzia presidente} "{president, France president}", the head word lemma of the mention candidate *presidente* is equal to a lemma in the set of enriched lemmas of the potential antecedent *presidente* and all the enriched lemmas of the potential antecedent appear in the cluster lemmas of the mention candidate, so the second constraint is fulfilled. This constraint aims to link coreferent mentions where a mention with novel information appears later in text than the less informative one. As pointed out in Fox (1993), it is not common to introduce novel information in later mentions but it sometimes happens.

4.3 Synonymy sieve

To create this new sieve, we have extracted from Basque WordNet (Pociello et al., 2011) all the words that are considered synonyms in this ontology. The Basque WordNet contains 32,456 synsets and 26,565 lemmas, and is complemented by a hand-tagged corpus comprising 59,968 annotations (Pociello et al., 2011).

From all synsets, a static list of 16,771 sets of synonyms has been created and integrated in the coreference resolution system. Using the synonyms' static list, the *Synonymy sieve* considers two mentions as coreferent if the following constraints are fulfilled: i) the head word of the potential antecedent and the head word of the mention candidate are synonyms and ii) all the lemmas in the mention candidate's span are in the potential antecedent cluster word lemmas or *vice versa*. For example, the mention candidate *Libanoko legebiltzarra* "Lebanon parliament" and the *Libanoko parlametua* "Lebanon parliament" are considered coreferent as the head words *legebiltzarra* and *parlamentua* are synonyms and the lemma *Libano* "Lebanon" of the word *Libanoko* is present in the cluster word lemmas of the potential antecedent.

5 System evaluation

In order to quantify the impact of using semantic knowledge sources in coreference resolution, we have tested the enriched coreference resolution system using the EPEC corpus and compared the results with the baseline system. The experimentation has been carried out using automatic mentions and gold mentions. In both cases named entity disambiguation and entity linking has been performed automatically.

5.1 Metrics

The metrics used to evaluate the systems' performances are MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF$_e$ (Luo, 2005), CEAF$_m$ (Luo, 2005), BLANC (Recasens and Hovy, 2011) and LEA (Moosavi and Strube, 2016). The CoNLL metrics is the arithmetic mean of MUC, B^3 and CEAF$_e$ metrics. The scores have been calculated using the reference implementation of the CoNLL scorer (Pradhan et al., 2014).

5.2 Experimental results

As pointed out in Rahman and Ng (2011), while different knowledge sources have been shown to be useful when applied in isolation to a coreference system, it is also interesting to observe if they offer complementary benefits and can therefore further improve a resolver when applied in combination. In order to quantify the individual improvement of each new sieve, we compared the baseline system (1) with the system in which the wiki-alias sieve has been added (2), with the one where the synonymy sieve has been added (3), and with the final system combining both sieves (4).

Table 3 shows the results obtained by the baseline system compared with those obtained by the coreference resolution system, which uses semantic knowledge sources. These scores are obtained with automatically detected mentions (F_1 =77.57).

The scores obtained by systems using the gold mentions (F_1 =100), i.e., when providing all the correct mentions to the coreference resolution systems, are shown in Table 4.

6 Discussion

Observing the results presented in Table 3, we can see that the baseline system's F_1 scores are outperformed in all the metrics by the semantically enriched system. In CoNLL metric, the improved system has a score of 55.81, which is slightly higher than the baseline system, to be precise, 0.24 higher.

As shown in Table 4, the baseline F_1 scores are also outperformed in all the metrics, except in B^3 when gold mentions are used. The official CoNLL metric is improved by 0.39 points.

Regarding recall and precision scores when automatic and gold mentions are used, all the metrics

	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
1	34.1	55.76	42.32	57.98	68.83	62.94	60.78	62.31	61.54	66.02	58.41	61.98	38.41	53.57	43.18	46.71	51.78	49.12	55.74
2	34.41	55.70	42.54	58.09	68.64	62.93	60.73	62.26	61.49	65.94	58.49	61.99	38.65	53.27	43.35	46.82	51.64	49.11	55.82
3	34.57	56.03	42.76	58.08	68.80	62.98	60.85	62.38	61.61	65.99	58.51	62.03	38.53	53.65	43.31	46.83	51.97	49.27	55.92
4	34.88	55.90	**42.95***	58.19	68.60	**62.97**	60.80	62.33	**61.56**	65.92	58.60	**62.04**	38.77	53.33	**43.48***	46.94	51.83	**49.26**	**55.98***

Table 3: Results obtained when automatic mentions are used. 1=Baseline, 2=1+Wiki sieve, 3=2+Synonymy sieve, 4=1+Wiki sieve+Synonymy sieve.

	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
1	48.76	71.94	58.12	81.35	93.47	86.99	80.57	80.57	80.57	89.00	78.24	83.27	67.09	84.65	72.77	66.36	71.11	68.66	76.12
2	49.84	70.81	58.50	81.71	92.83	86.92	80.57	80.57	80.57	88.69	78.77	83.44	67.51	83.27	72.84	66.60	71.01	68.73	76.28
3	50.00	71.50	58.85	81.69	93.19	87.06	80.80	80.80	80.80	88.90	78.82	83.56	67.39	84.23	72.95	66.68	71.52	69.02	76.49
4	50.46	70.99	**58.99***	81.86	92.81	86.99	80.71	80.71	**80.71**	88.71	79.00	**83.57***	67.68	83.34	**73.00**	66.79	71.29	**68.97**	**76.51***

Table 4: Results obtained when gold mentions are used. 1=Baseline, 2=1+Wiki sieve, 3=2+Synonymy sieve, 4=1+Wiki sieve+Synonymy sieve.

except $CEAF_e$ show an improvement in recall and decrease in precision when two new sieves are applied. The reason why the $CEAF_e$ metric is behaving differently could be that, as mentioned by Denis and Baldridge (2009), CEAF ignores all correct decisions of unaligned response entities. Consequently, the CEAF metric may lead to unreliable results.

It is interesting to compare the improvements obtained by the system which uses semantic knowledge sources in CoNLL scores. The improvement when automatic mentions are used is lower than when gold mentions are provided, 0.24 and 0.39 respectively. In both cases, even the improvements obtained are modest, they are statistically significant using Paired Student's t-test with p-value < 0.05.

As pointed out in Versley et al. (2016), in realistic settings, where the loss in precision would be amplified by the additional non-gold mentions, it is substantially harder to achieve gains by incorporate lexical and encyclopedic knowledge, but possible and necessary. A similar idea is concluded by Durrett and Klein (2013). They mention that despite the fact that absolute performance numbers are much higher on gold mentions and there is less room for improvement, the semantic features help much more than they do in system mentions.

To conclude the analysis of the results, it is also interesting to observe the difference between the results obtained by both systems when automatic mentions and when gold mentions are used. It is clear that having accurate preprocessing tools and a good mention detector are crucial to obtain good results in coreference resolution. In both sys-

tems the difference in CoNLL score is about 20.00 points higher when gold mentions are used.

The results obtained have enabled us to carry out a new error analysis in the development set. After applying the new two sieves, the error analysis has revealed four major issues that directly affect not obtaining bigger improvement when knowledge resources are used:

1. Some mentions do not have Wikipedia entry, as the coverage of Basque Wikipedia (257,546 pages) has less coverage than other languages, for example English (5,250,837 pages), i.e., Basque version is 21 times smaller.

2. Due to incorrect mention disambiguation, some mentions are linked to incorrect Wikipedia pages. The precision obtained in disambiguation is 87,84%.

3. Precision errors, provoked by cases where many proper noun mentions were potential antecedent for a common noun. For example, *Oslo* is linked by *hiriburu* "capital", nevertheless the correct antecedent for *hiriburu* is another capital that appears in text, in this specific case, *Jerusalem*.

4. Some indefinite mentions which do not have antecedent are linked incorrectly. For example, *estaturik* "state" is linked with *Frantziak* "France".

5. In the synonyms' static list, some synonyms that appear in texts are missing. In addition, many synonyms are so generic, i.e., they

are synonyms depending on the context in which they appear. As a consequence of missing synonyms, some mentions with synonymy relations between them are not linked. The presence of very generic synonyms provokes to incorrectly link mentions that are not coreferent, so that precision decreases. Identifying the particular sense that a word has in context would likely help to improve the precision.

Regarding the issues that affect improvement of the systems when knowledge bases are used, Uryupina et al. (2011) suggest that in their particular case the errors introduced are not caused by any deficiencies in web knowledge bases, but reflect the complex nature of the coreference resolution task.

7 Conclusions and future work

We have enriched the Basque coreference resolution adding new two sieves, *Wiki-alias* and *Synonymy sieve*, respectively. The first sieve uses the enriched information of named-entity mentions after they have been linked to their correspondent Wikipedia page, using Entity Linking techniques. The second sieve uses a static list of synonyms extracted from Basque WordNet to consider whether two mentions are coreferent.

Applying the two new sieves, the system obtains an improvement of 0.24 points in CoNLL F_1 when automatic mentions are used and the CoNLL score is outperformed by 0.39 points when the gold mentions are provided. The error analysis of the enriched system has revealed that the knowledge bases used, Basque Wikipedia and Basque WordNet, have deficiencies in their coverage compared with knowledge bases in major languages, for example, English. We suggest that there is margin of improvement, as Basque Wikipedia and Basque WordNet coverage increase, bearing in mind that coreference resolution is a complex task.

As future work, we intend to improve the Pronoun resolution and Ellipsis Resolution, as we observed in the error analysis presented in Section 2 they are the cause of considerable coreference resolution errors, around % 12 of total errors.

Acknowledgments

References

Itziar Aduriz, María Jesús Aranzabe, Jose Mari Arriola, Maite Atutxa, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, and Ruben Urizar. 2006. Methodology and Steps towards the Construction of EPEC, a Corpus of Written Basque Tagged at Morphological and Syntactic Levels for the Automatic Processing. In *Language and Computers*, volume 56, pages 1–15. Rodopi, Amsterdam, Netherlands.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Athens, Greece. Association for Computational Linguistics.

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random Walks for Knowledge-based Word Sense Disambiguation. *Comput. Linguist.*, 40(1):57–84.

Amit Bagga and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, Granada, Spain.

Klara Ceberio, Itziar Aduriz, Arantza Díaz de Ilarraza, and Ines Garcia-Azkoaga. 2016. Coreferential relations in Basque: the annotation process. In *Theoretical Developments in Hispanic Linguistics*. The Ohio State University.

Pascal Denis and Jason Baldridge. 2009. Global Joint Models for Coreference Resolution and Named Entity Classification. *Procesamiento del Lenguaje Natural*, 43:87–96.

Greg Durrett and Dan Klein. 2013. Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2014. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *TACL*, 2:477–490.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Barbara A. Fox. 1993. *Discourse structure and anaphora: written and conversational English*. Cambridge Univeristy Press.

Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke Zettlemoyer. 2013. Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 289–299, Seattle, Washington, USA. Association for Computational Linguistics.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-Driven Analysis of Challenges in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Computational Linguistics*, 39(4):885–916.

Xiaoqiang Luo. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Pablo Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia: A Multilingual Cross-domain Knowledge Base. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Nafise Sadat Moosavi and Michael Strube. 2016. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2):121–142.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 192–199. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2011. Coreference Resolution with World Knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 814–824, Portland, Oregon. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2012. Learning-based Multi-sieve Co-reference Resolution with Knowledge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1234–1244, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Marta Recasens, Matthew Can, and Daniel Jurafsky. 2013. Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 897–906, Atlanta, Georgia. Association for Computational Linguistics.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Arantza Díaz de Ilarraza. 2015. Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. *Procesamiento del Lenguaje Natural*, 55:23–30.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Arantza Díaz De Ilarraza. 2016. Improving mention detection for Basque based on a deep error analysis. *Natural Language Engineering*, pages 1–34, 007.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, pages 697–706.

Olga Uryupina, Massimo Poesio, Claudio Giuliano, and Kateryna Tymoshenko. 2011. Disambiguation and Filtering Methods in Using Web Knowledge for Coreference Resolution. In *FLAIRS Conference*, pages 317–322. AAAI Press.

Yannick Versley, Massimo Poesio, and Simone Ponzetto. 2016. Using Lexical and Encyclopedic Knowledge. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora Resolution: Algorithms, Resources, and Applications*, pages 393–429. Springer Berlin Heidelberg, Berlin, Heidelberg.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 45–52. Association for Computational Linguistics.