Natural Language Engineering **23** (3): 351–384. © Cambridge University Press 2016
doi:10.1017/S1351324916000206

Improving mention detection for Basque based on a deep error analysis

ANDER SORALUZE, OLATZ ARREGI, XABIER ARREGI and ARANTZA DÍAZ DE ILARRAZA

IXA Group, University of the Basque Country, Donostia-San Sebastián, Spain.
e-mails ander.soraluze@ehu.es, olatz.arregi@ehu.es, xabier.arregi@ehu.es, a.diazdeilarraza@ehu.es.

(*Received 22 January 2015; revised 13 May 2016; accepted 16 May 2016; first published online 12 July 2016*)

Abstract

This paper presents the improvement process of a mention detector for Basque. The system is rule-based and takes into account the characteristics of mentions in Basque. A classification of error types is proposed based on the errors that occur during mention detection. A deep error analysis distinguishing error types and causes is presented and improvements are proposed. At the final stage, the system obtains an F-measure of 74.57% under the Exact Matching protocol and of 80.57% under Lenient Matching. We also show the performance of the mention detector with gold standard data as input, in order to omit errors caused by the previous stages of linguistic processing. In this scenario, we obtain an F-measure of 85.89% with Strict Matching and of 89.06% with Lenient Matching, i.e., a difference of 11.32 and 8.49 percentage points, respectively. Finally, how improvements in mention detection affect coreference resolution is analysed.

1 Introduction

Coreference resolution consists of identifying textual expressions (mentions) that refer to real-world objects (entities) and of determining whether these mentions refer to the same entity. It is well known that coreference resolution is essential in Natural Language Processing (NLP) applications where a higher accuracy in discourse analysis leads to better performance. Some of the tasks that can benefit from coreference resolution include information extraction, question answering, machine translation, sentiment analysis, machine reading, text summarisation, and text simplification.

In Soraluze *et al.* (2012), we presented a preliminary mention detector for Basque, evaluated on a small corpus. In this paper, we present an improved version that has been evaluated on a considerably larger corpus to obtain more reliable results. A bigger corpus has also made it possible to carry out a deep error analysis, which in turn has allowed us to identify and address some of the main sources of error.

Basque is a non-Indo-European language and differs considerably in grammar from the languages spoken in the surrounding regions. It is an agglutinative,

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

head-final, pro-drop, free-word order language isolate (Laka 1996), which presents a number of special problems regarding mention detection.

The paper is structured as follows. After reviewing related work in Section 2, we describe the types of mention structures in Basque in Section 3. Section 4 compares mention types in Basque, English, Catalan, and Spanish. Section 5 describes the experiments we carried out in order to create a mention detector, and in Section 6, a classification of errors types and causes made during mention detection is presented. Moreover, after a deep error analysis is carried out, further improvements are designed. Section 7 outlines an experiment with running the detector on gold input, with the aim of observing the effect of general natural language preprocessing tools. In order to observe how improvements in mention detection affect coreference resolution, in Section 8, we compare the scores obtained by a coreference model before and after the improvements in mention detection are applied. Finally, conclusions and future work are discussed.

2 Related work

Much attention has been paid to the problem of coreference resolution in the past two decades. Conferences specifically focusing on coreference resolution have been organised since 1995. The sixth and seventh Message Understanding Conferences (MUC-6 1995; MUC-7 1998) included a specific task on coreference resolution. The Automatic Context Extraction (ACE) Program focused on identifying certain types of relations between a predefined set of entities (Doddington *et al.* 2004) while the Anaphora Resolution Exercise (ARE) involved anaphora resolution and NP coreference resolution (Orasan *et al.* 2008).

More recently, SemEval-2010 Task 1 was dedicated to coreference resolution in multiple languages (Recasens *et al.* 2010). A year later, in the CoNLL-2011 shared task (Pradhan *et al.* 2011), participants had to model unrestricted coreference in the English-language OntoNotes corpora (Pradhan *et al.* 2007) and CoNLL 2012 Shared Task (Pradhan *et al.* 2012) involved predicting coreference in three languages: English, Chinese, and Arabic.

The coreference resolution task is commonly divided into two main subtasks: mention detection and resolution of references (Pradhan *et al.* 2011). Mention detection is considered a crucial task for the accuracy of end-to-end coreference resolution systems and the mention detection step has a significant impact on coreference resolution systems: Uryupina (2008) reports that 35% of recall errors in their coreference resolution system are caused by missing mentions and in Uryupina (2010) adds that 20% of precision errors are due to inaccurate mention detection, while Stoyanov *et al.* (2009) conclude that improving the ability of coreference resolvers to identify coreference elements or mentions would likely improve the state-of-the-art greatly. Chang *et al.* (2011) argue that a robust mention detector is crucial, as detection errors propagate to the coreference stage; as they show, their system, which uses gold mentions, outperforms the system that uses predicted or system mentions by a large margin (from 15% to 18% absolute difference in F_1 score). Hacioglu, Douglas and Chen (2005) also show that errors in the first stages,

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

such as mention detection, propagate, and reduce the performance of subsequent stages and Zhekova and Kübler (2010) demonstrated experimentally that during the development phase the detection of mentions is one of the most important steps to achieve a high accuracy in coreference resolution. Also, in Broscheit *et al.* (2010) the importance of mention detection is recognised.

There are two main approaches to the technology used to create mention detectors: rule-based systems and machine learning models. Machine learning models seem able to better balance precision and recall and thus to achieve higher F-score values in the mention detection task, their recall tends to be quiet a bit lower than obtained by rule-based systems designed to favour recall (Pradhan *et al.* 2011). In end-to-end coreference resolution systems, low recall in the mention detection step has negative effects on the whole process, since missed mentions are not available for coreference resolution in the later stages. As is pointed out in Uryupina and Moschitti (2013), in the CoNLL 2011 shared task the majority of the participants relied on rule-based modules to obtain mention boundaries for English. However, in the CoNLL 2012 shared task, participants with rule-based systems fell back to very simple baselines in mention detection, which demonstrated that rule-based systems are not easily adaptable to languages other than the one for which they were created.

The first systems that resolved coreferences were mainly designed for English. During the past few years, however, efforts in resolving coreference in languages other than English have been published. For example, in the SemEval-2010 Task 1, Catalan, Dutch, English, German, and Italian coreference resolvers were evaluated. The CoNLL 2012 shared task added Arabic and Chinese. Less-resourced languages are also gaining presence in coreference resolution, as exemplified by recent work on Hungarian (Mihăltz 2008), Polish (Ogrodniczuk and Kopeć 2011), Czech (Nguy, Novák and Żabokrtský 2009), and also Indian Languages (Lalitha Devi, Sundar Ram and RK Rao 2014).

3 Mentions in Basque

Language-specific patterns that vary according to the features of each language have to be taken into consideration during mention detection. In Basque written text, while not all mentions represent a particular detection challenge, some can have complex structures and require a more sophisticated identification strategy.

In general, we consider noun phrases (NP), focusing on the largest span of the NP. In the case of nouns complemented by subordinate clauses and coordination, NPs embedded in a larger NP are also extracted.

These are the structures that we consider mentions:

- **Proper Nouns (PN):** Structures that have a proper noun as head.

  (1) [Clinton] itxaropentsu agertu zen kazetarien aurrean.

  Clinton itxaropentsu agertu_zen kazetarien aurrean
  Clinton hopefully appeared the_reporters in_front_of

  "[Clinton] appeared hopeful in front of the reporters."

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

- **Pronouns (Pro):** All personal pronouns are considered mentions. However, in Basque, the demonstrative determiners also act as third person pronouns (Laka 1996). To deal with this ambiguity, we mark as mentions the demonstratives used as pronouns (example 2).

  (2) LDPko buruek Mori hautatu zuten apirilean Keizo Obuchi lehen ministroa ordezkatzeko, [hark] tronbosia izan ostean.

  LDPko buruek Mori hautatu_zuten apirilean Keizo Obuchi lehen
  of_LDP The_heads Mori chose in_April Keizo Obuchi Prime
  ministroa ordezkatzeko hark tronbosia izan ostean
  the_Minister to_replace who_(he) a_thrombosis suffered after

  "The heads of LDP chose Mori in April to replace the Prime Minister Keizo Obuchi after [who (he)] suffered a thrombosis"

  Note that "hark" acts as a pronoun in this context, but the same word may be used as a determiner in other contexts. For instance, "hark" is a determiner in the sentence "Gizon [hark] tronbosia izan zuen;" ([That] man suffered a thrombosis).
- **Possessives (Poss):** We consider two types of possessives: possessive determiners, even if they are not the head of the NP as in example 3; and possessive pronouns, as in example 4.

  (3) Ekisoainek Granollersera joatea nahiago du, [bere] emazteak lagunak baititu bertan.

  Ekisoainek Granollersera joatea nahiago_du bere emazteak lagunak
  Ekiosain to_Granollers to_go prefers his wife friends
  baititu bertan
  has_because there

  "Ekisoain prefers to go to Granollers because [his] wife has friends there."

  (4) Escuderok euskal musika tradizionala eraberritu eta indartu zuen. [Harenak] dira, esate baterako, Illeta, Pinceladas Vascas eta Eusko Salmoa obrak.

  Escuderok euskal musika tradizionala eraberritu eta
  Escudero Basque music traditional renewed and
  indartu_zuen [Harenak] dira esate_baterako Illeta Pinceladas
  gave_prominence his are for_example Illeta Pinceladas
  Vascas eta Eusko Salmoa obrak
  Vascas and Eusko Salmoa The_works

  "Escudero renewed and gave prominence to traditional Basque music. The works Illeta, Pinceladas Vascas, and Eusko Salmoa, for example, are [his]."

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

- **Verbal nouns (VN):** Verbs that have been nominalised and function as the head of the mention with the corresponding case marking suffix. The whole clause governed by the verbal noun has to be annotated.

  (5) [Europar Batasunaren zabaltze honek] arazo asko konpontzera behartuko ditu.

  Europar Batasunaren zabaltze_honek arazo asko
  European of_the_Community the_growth problems many
  konpontzera behartuko_ditu
  to_solve will_force_them

  "[The growth of the European Community] will force them to solve many problems."

- **NPs as part of complex postpositions (CPost):** As Basque has a postpositional system, we mark the independent NP that precedes the complex postpositions. In example 6, the postposition is *aurka* (against) and we annotate the noun Athleticen (in this case, a proper noun) that precedes it, which is in genitive case (-en).

  (6) Moreno eta Vlatko Djolonga [Athleticen] aurka jokatzeko moduan daude.

  Moreno eta Vlatko Djolonga Athleticen aurka jokatzeko moduan
  Moreno and Vlatko Djolonga Athletic against to_play in_shape
  daude.
  are

  "Moreno and Vlatko Djolonga are in shape to play against [Athletic]."

- **NPs containing subordinate clauses (SubrCl):** The head of these mentions is always a noun complemented by a subordinate clause. In example 7, the head noun is complemented by a subordinate clause that is called a complementary clause. We take the whole span of the NP (both the subordinate clause and the head noun) as a mention. In addition, when a relative clause adds information to the noun, as in example 8, the boundaries of the mention are set from the beginning of the relative clause to the end of the NP.

  (7) [Oslon hasitako prozesua] gaur bukatuko da.

  Oslon hasitako prozesua gaur bukatuko_da
  in_Oslo that_began The_process today is_going_to_end

  "[The process that began in Oslo] is going to end today."

  (8) [Antimisilen inguruan Pentagonoa atontzen ari den sistema] aurkeztu dute.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Antimisilen_inguruan Pentagonoa atontzen ari_den sistema
The_missile_defense the_Pentagon preparing that_is system
aurkeztu_dute
has_been_presented

"[The missile defense system that the Pentagon is preparing] has been presented."

- **Ellipsis (Ellip):** In Basque, ellipsis is a common phenomenon. At morphosyntactic level, a noun-ellipsis occurs when the suffixes attached to the word correspond to a noun, even when the noun is not explicit in the word. We consider this type of ellipsis in the case of verbs that take suffixes indicating noun-ellipsis, as in example 9. The POS given by the analyser indicates the presence of an ellipsis, which is deduced by the presence of both the verb (*jaitsi ginen-* we were relegated) and the empty mark (*-Ø-ekoa* that in which). All the information corresponding to both units is stored and treated as a noun.

  (9) Niretzat oso urte txarra izan zen [bigarren mailara jaitsi ginenekoa].

  Niretzat oso urte txarra izan_zen bigarren mailara
  In_my_opinion really year a_bad it_was second to_the_division
  jaitsi_ginenekoa
  that_in_which_we_were_relegated

  "In my opinion it was a really bad year, [that in which we were relegated to the second division]."

- **Coordination (Coor):** In the case of coordination, nominal groups of a conjoined NP are extracted. We also regard as mentions the nested NPs (*Xabier Mikel Errekondo* and *Alvaro Jauregi*) and the whole coordinate structure (*Xabier Mikel Errekondo eta Alvaro Jauregi*).

  (10) [[Xabier Mikel Errekondo] eta [Alvaro Jauregi]] euren kontratuak berritzear daude.

  Xabier Mikel Errekondo eta Alvaro Jauregi euren kontratuak
  Xabier Mikel Errekondo and Alvaro Jauregi their contracts
  berritzear daude
  about_to_renew are

  "[[Xabier Mikel Errekondo] and [Alvaro Jauregi]] are about to renew their contracts."

- **Location Adverbs (LocAdv):** In general, adverbs are not referential, yet location adverbs do have a referential function. Therefore, location adverbs are considered mentions.

  (11) Futbol jokalariak Biarritzera joatekoak ziren, [han] festa antolatuta baitzuten.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 1. *Mention types in EPEC corpus*

Mention type	#	%
Proper nouns	3047	23.82
Pronouns	379	2.97
Possessives	174	1.36
Verbal nouns	289	2.26
NPs as part of complex postpositions	599	4.68
NPs containing subordinate clauses	403	3.15
Ellipsis	78	0.61
Coordination	367	2.87
Location adverbs	68	0.53
Common noun phrases	7388	57.75
Total	12, 792	100.00

Futbol jokalariak Biarritzera joatekoak_ziren han festa
football The_players to_Biarritz intented_to_go there a_party
antolatuta baitzuten
organised because_they_had

"The football players intended to go to Biarritz, because they had a party organised [there]."

- **Common Noun Phrases (CNP):** Phrases that have a noun as head word.

(12) [Langileak] haserre daude hartutako erabakiarekin.

Langileak haserre daude hartutako erabakiarekin
The_workers angry are taken with_the_decision

"[The workers] are angry with the decision taken."

In Table 1 we can observe the distribution of mention types in Basque according to the classification of mentions presented above. This distribution has been calculated on the EPEC corpus (the Reference Corpus for the Processing of Basque), which has been previously annotated manually. It is a collection of news published in *Euskaldunon Egunkaria*, a Basque newspaper. More details about this corpus are presented in Section 5.

As can be observed, the majority of mentions are common noun phrases, (57.75%), followed by proper nouns (23.82%). The other structures are much less frequent in the corpus, ranging from about 0.5% to just under 5% of the corpus.

4 Comparison of mention types between different languages

This section compares the distribution of mentions in Basque with their distribution in English, Catalan, and Spanish.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 2. *Comparison of mention types among different languages: Basque (Eu), English (En), Spanish (Es), and Catalan (Cat)*

Mention type	Eu	En	Es	Cat
Proper nouns	23.82	2.72	13.5	12.43
Pronouns	2.97	25.85	14.65	14.24
Possessives	1.36	8.78	3.55	3.12
Verbal nouns	2.26	1.93	–	–
NPs as part of complex postpositions	4.68	–	–	–
NPs containing subordinate clauses	3.15	–	–	–
Ellipsis	0.61	–	–	–
Coordination	2.87	–	3.41	3.97
Location adverbs	0.53	–	–	–
Common noun phrases	57.75	58.85	60.62	62.86

Pradhan *et al.* (2011) present a distribution of mentions in OntoNotes corpus by their syntactic category. This English corpus collects newswire articles, magazine articles, broadcast news, broadcast conversations, and web data. Recasens and Martí (2010) present the statistics of mention types in the AnCora corpus for Catalan and Spanish. Each corpus consists of newspaper and newswire articles.

The distributions for English, Catalan, and Spanish are not directly comparable with the distribution for Basque. For instance, the type "NPs as part of complex postpositions" has no equivalent type in English, Spanish or Catalan. To make them partially comparable, we strove to map the mention types we consider in Basque to the same or similar structures in these other languages.

Although the comparison is not detailed, it enables us to reach certain conclusions about the differences between Basque and other languages.

Table 2 presents the distributions in different languages. Some differences and similarities between languages can be noted. As expected, in all the languages the majority of mention structures are common noun phrases, with quite similar percentages: 57.75% in Basque, 58.85% in English, 60.62% in Spanish, and 62.86% in Catalan. In Basque, proper nouns also appear frequently, 23.82%; this does not happen in English, only 2.72%. In Spanish and Catalan, the presence of proper nouns is similar, 13.5% and 12.43%, respectively; however, they are fewer than in English and more numerous than in Basque. Another remarkable difference occurs regarding the presence of pronouns. Again, the values are similar for Catalan and Spanish, but there is a significant difference between English and Basque: in English, pronouns are far more frequent than in Basque. As pointed out in Recasens and Hovy (2010), AnCora has a smaller number of personal pronouns compared with English because Spanish and Catalan are pro-drop languages, which allow zero subject pronouns that can be inferred from the verb. In addition to the pronominal subject, Basque allows the omission of direct and indirect objects (three-way pro-drop) and, contrary to English, Catalan, and Spanish, it does not generally use pronouns to construct relative clauses. These characteristics can be the reason for

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

the low presence of pronouns in Basque compared with English, Catalan, and Spanish.

The other differences are far less dramatic. The main outlier is possessive structures in English, which are clearly more frequent than in the other languages.

After reviewing the presence of mention types in different languages, some general conclusions can be sketched out. In particular, it seems that the sum of proper nouns and pronouns is about the same in each language: 26.79% for Basque, 28.57% for English, 28.15% for Spanish, and 26.67% for Catalan. This may suggest that each language behaves differently with regard to the use of pronouns. While English tends to use pronouns to reference mentions that have appeared previously, Spanish and Catalan balance the use of pronouns and proper nouns, using both almost equally. Pronouns have no gender in Basque, which seems to prioritise the use of proper nouns instead of pronouns, presumably to avoid ambiguity.

5 Experimental setup

5.1 Corpus

The EPEC corpus (the Reference Corpus for the Processing of Basque) aims to function as a "reference" corpus for the development and improvement of several NLP tools for Basque (Aduriz *et al.* 2006). It is a 300,000-word sample collection of news published in *Euskaldunon Egunkaria*, a Basque-language newspaper. This corpus has been manually tagged at different levels (morphology, syntax, phrases, etc.). Recently, mentions and coreference chains have also been tagged by two expert linguists in a subpart of EPEC corpus. The estimation of the inter-annotator agreement, including the chance-based factor for the task of tagging mentions has not been researched yet according to Artstein and Poesio (2008). In Kopeć and Ogrodniczuk (2014) only the observed agreement is presented, since it is difficult to estimate the probability of a random tagging of a mention. The same procedure is used in Ohta *et al.* (2012). One annotation is considered as gold and the other as system. We evaluated the annotation consistency in mention detection in the same way, which was scored using the Strict Matching protocol, explained in Section 5.2. The F-measure obtained for agreement was 94.07%.

The mention detector has been developed and tested using a subpart of the EPEC corpus consisting of 46,383 words that correspond to 12,792 mentions. The relation between the number of words and the number of mentions is about 27%, meaning that approximately every four words a mention is detected. Similar statistics regarding the proportion of words/mentions can be found in Màrquez, Recasens, and Sapena (2013), where English, Spanish, and Catalan values are presented.

5.2 Scoring protocols

The most used measures to evaluate mention detection are precision, recall, and F-measure. To calculate these, mentions extracted by the mention detector and the set of manually tagged mentions are compared. Typically, two matching methods have been used to make the comparison between mentions.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

The first one, known as *Lenient Matching* or *Partial Matching*, considers mentions to be correct if their span is within the span of the gold mention and contains the head word (Kummerfeld *et al.* 2011). This protocol was used in Message Understanding Conferences (MUC-6 1995; MUC-7 1998). Slightly different measure was used in SemEval-2010 Task 1 (Recasens *et al.* 2010) to score the mention detection task. Mention were rewarded with one point if their boundaries coincide with those of gold mentions and the mentions that whose boundaries are within the gold mention including its head with 0.5 points. And in ACE programs (NIST 2008) only system mentions that overlap more than 30% in terms of character span with the gold mention and which have the same entity type were credited.

The second scoring protocol is clearly a stricter variation, because it only considers correct mentions that are exactly the same as the gold mentions. This matching method is known as *Strict Matching* and has been used in conferences such as CoNLL-2011 Shared Task (Pradhan *et al.* 2011).

We present our mention detector scores using both Lenient Matching and Strict Matching.

5.3 Nominal chunks as mentions

For mention detection, as a starting point we decided to use a rule-based approach using a generic NLP tool, namely a chunker (Aduriz and Díaz de Ilarraza 2003), an analyser that identifies verbal and nominal chunks based on rule-based grammars. This tool is integrated in a pipeline for Basque processing where other modules are also applied: (i) A morphological analyser that performs word segmentation and PoS tagging (Alegria *et al.* 1996), (ii) A lemmatiser that also disambiguates the PoS and the syntactic function (Alegria *et al.* 2002), (iii) A multi-word item identifier that determines which groups of two or more words are to be considered multi-word expressions (Alegria *et al.* 2004), (iv) A named-entity recogniser that identifies and classifies named entities (person, organization, location) in the text (Alegria *et al.* 2003).

The chunker uses Constraint Grammar (CG) formalism (Karlsson *et al.* 1995) and obtains a precision value of 81.08%, a recall of 81.09% and F-measure of 81.08% in chunking.

The nominal-chunks can be considered the structures that are most similar to mentions, so we consider them and discard the verbal chunks for mention detection. The obtained results can be seen in Table 3. We obtained an F-measure of 64.28% using the Strict Matching protocol and 71.12% using Lenient Matching. These scores were considered as baseline.

Regarding the scores obtained using only a chunker for mention detection, we conclude that mention detection itself is a very challenging task since expressions can have complex syntactic and semantic structures. Preprocessing with generic NLP tools, while helpful, did not always succeed in identifying mention boundaries correctly. The same fact is pointed out by Nguyen, Kim and Tsujii (2008), who suggest that using a base noun phrase chunker is insufficient for mention detection in the bio-domain, and that therefore some adaptations are needed. In less resourced

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 3. *Mention detection using nominal chunks. B=Baseline*

	Strict matching			Lenient matching		
	P	R	F_1	P	R	F_1
B	62.24	66.46	64.28	67.02	75.75	71.12

	[Mikelek]	erosi	zituen	[etxeak]	...
Chunker:	NP	BVP	EVP	NP	
Correct:	[[Mikelek]	erosi	zituen	etxeak]	...

Fig. 1. Wrongly identified mention. NP =Noun Phrase, BVP = Begin Verbal Phrase, EVP = End Verbal Phase.

languages such as Basque, a processing of the output obtained by generic NLP tools is needed to obtain better results in mention detection. In contrast, in richer resource languages where generic NLP tools such as parsers obtain good results, these tools can be used for accurate mention detection with little or no processing.

The example in Figure 1 illustrates the differences between the chunker output and mentions. In this example, we deal with the sentence *Mikelek erosi zituen etxeak*... "The houses that Mikel bought..". The first row shows that *Mikel* "Mikel" and *etxeak* "houses" are the nominal chunks obtained by the chunker, which are automatically considered mentions in this first approach. *Mikel* is a mention itself but *etxeak* cannot be considered a mention, because it is only the head of a larger mention *Mikelek erosi zituen etxeak* "The houses that Mikel bought". It is clear that some processing of the output generated by a chunker is needed to correctly identify mentions.

5.4 Mention detection with finite state transducers

As explained in Section 5.3, using a preprocessing tool such a chunker does not really succeed in mention identification. Some mention structures are more complex than nominal chunks and mentions exceed the chunk boundaries. Thus, to identify even the complex type of mentions, we decided to combine the chunker with a clause tagger, that is, an analyser that identifies clauses, combining rule-based-grammars and machine learning techniques (Arrieta 2010).

We combined the tags provided by these tools and created a mention detector system, defining a set of hand-crafted rules that have been compiled into Finite State Transducers (FST).

The FSTs are able to detect complex structures that should be identified as mentions and that chunkers do not recognise. In reference to the classification of mentions presented in Section 3, the FSTs detect verbal nouns, NPs as part of complex postpositions, NPs containing subordinate clauses, ellipses, coordination cases, and location adverbs.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Input	[Armada	britainiarrak]	[Ipar	Irlandan]	dituen	[bi	kuartel]	eraitsi	dituzte
Chunker	BNP	ENP	BNP	ENP	VREL	BNP	ENP	BVP	EVP
Clause tagger	{CB{CB				CB}				CB}
Output	[[Armada	britainiarrak]	[Ipar	Irlandan]	dituen	bi	[kuartel]]	eraitsi	dituzte

Fig. 2. The input and the output of the mention detector. BNP = Begin-NP, ENP = End-NP, BVP = Begin-VP, EVP = End-VP, VREL = Relative Verb, CB = Clause Boundary.

```
define RV Verb & $['VREL'];
define RM [CB W+ RV NP [[and|or|...] NP]*] @-> "<MENTION>" ... "</MENTION>";
```

Fig. 3. A simplified rule to recognise NPs containing subordinate clauses. NP = Noun Phrase, CB = Clause Boundary, W=Word, RV = Relative Verb, RM = Relative Mention.

To better understand how the FSTs work, an example of a mention composed by an NP containing a relative clause is presented, and then a rule that identifies these structures and marks them as mentions is defined.

Suppose that the mention detector receives as input the sentence *Armada britainiarrak Ipar Irlandan dituen bi kuartel eraitsi dituzte*.

> *Armada britainiarrak Ipar Irlandan dituen bi*
> army British Northern in_Ireland that_the_has two
> *kuartel eraitsi_dituzte*
> military_barracks has_been_demolished
> "Two military barracks that the British army has in Northern Ireland have been demolished"

Figure 2 shows the above sentence with information obtained during the preprocessing step and the mentions that the system proposes.

Using the preprocessing information, the rule in Figure 3 identifies the verb containing a relative suffix (*dituen* "that has"), which is tagged with VREL.Next, the right side boundary is established in the NP or coordinated NPs that follow the relative verb (*bi kuartel* Two military barracks). Finally, the left side boundary of the mention is established at the closest clause boundary ({CB) to the left (*Armada* army). Following these steps, the system proposes the correct mention (*Armada britainiarrak Ipar Irlandan dituen bi kuartel* Two military barracks that the British army has in Northern Ireland) and tags the whole structure to show where the mention begins (<MENTION>) and ends(</MENTION>).

Further discussion about the behaviour of the FSTs can be found in Soraluze *et al.* (2012).

The use of Finite State Technology enables the processing of large datasets at a high processing speed and with low memory usage. Using *foma*[1] (Hulden 2009), an open source platform for finite-state automata and transducers, we defined 12 FSTs, composed of 34 hand-crafted rules.

[1] The regular expression syntax used in *foma* can be consulted at http://code.google.com/p/foma/

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 4. *Baseline and mention detection scores. B=Baseline, BMD=Basque Mention Detector*

	Strict matching			Lenient matching		
	P	R	F_1	P	R	F_1
B	62.24	66.46	64.28	67.02	75.75	71.12
BMD	73.86	72.87	73.36	78.69	80.71	79.68

Table 5. *Comparison of different mention detectors. SMD=Stanford Mention Detector, BRTMD= BART Mention Detector, BMD=Basque Mention Detector*

	Strict matching			Lenient matching		
	P	R	F_1	P	R	F_1
SMD	23.15	29.22	25.83	50.25	65.98	57.05
BRTMD	34.69	42.87	38.35	54.89	72.27	62.39
BMD	73.86	72.87	73.36	78.69	80.71	79.68

We compared the results of the mention detector thus obtained with a baseline, established by considering all the nominal chunks that the chunker outputs to be mentions. The results are presented in Table 4.

Our mention detector outperforms the Strict Matching protocol F-measure baseline by 9.08 percentage points, and that of Lenient Matching by 8.56 percentage points. The improvements obtained by our system are significant.[2]

With the aim of comparing our mention detector system with standard systems, we carried out experiments with the Stanford Mention Detector (SMD) presented in Lee *et al.* (2013) and with BART (Versley *et al.* 2008) for mention detection. Using a high-recall algorithm, SMD identifies nominal and pronominal mentions from all NPs, pronouns, and named entity mentions, and then non-mentions are filtered out. BART uses a built-in mention extraction module, computing boundaries heuristically from the output of a parser. It creates a list of candidate mentions by merging basic NP chunks with named entities.

The tokens and PoS tags provided to both mention detectors were analysed by Basque Linguistic processors, nevertheless, the PoS tags were mapped to equivalent Penn PoS tags (Marcus, Marcinkiewicz and Santorini 1993). The parse trees used by the two systems were created using the Stanford PCFG parser (Klein and Manning 2003).

Table 5 presents the scores obtained by the three compared systems. SMD, BART Mention Detector (BRTMD) and Basque Mention Detector (BMD).

[2] Statistical significance is tested with Paired Student's t-test. p value < 0.01

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Compared to the SMD, our mention detector outperforms Strict Matching protocol F-measure by 47.53 and that of Lenient Matching by 22.63 percentage points. BMD also improves the scores that BRTMD obtains by 35.01 when Strict Matching protocol is used, and by 17.29 using Lenient Matching. The great difference in F-measures of Stanford and BRTMD compared with those obtained by our mention detector clearly shows that language specific characteristics (agglutinative, free-word order, etc. for Basque) should be taken into account when mention detection is performed.

6 Error analysis

Evaluation scores can show how efficient a mention detector is; however, they neither identify the type of errors that the mention detector makes, nor give any indication of how those errors might be corrected.

In Subsection 6.1, we present the classification of the error types we found during the error analysis, and in 6.2 the causes of these errors.

6.1 Error types

To help identify the different types of errors, we carried out a qualitative evaluation on a subcorpus of 1,904 mentions that were used to develop the system.

We classify the error types made in the mention detection step in eight different categories. The error categorization idea is based on (Kummerfeld and Klein 2013) where errors of a coreference system are classified. We have adapted this categorization to classify errors produced in mention detection. To better understand this, for each category we present an example. The first item in each example is the incorrect case (×) and the second the respective correct one (√). The categories are the following:

1. **Missing mention (MM):** The mention in gold standard has not been identified.

   × Guk bere garaian apustu handia egin genuen Sa Pinto fitxatu genuenean, jokalaria egoera berezi batean aurkitzen baitzen.

   √ Guk bere garaian apustu handia egin genuen Sa Pinto fitxatu genuenean, [jokalaria] egoera berezi batean aurkitzen baitzen.

Guk	*bere_garaian*	*apustu*	*handia*	*egin_genuen*	*Sa*	*Pinto*
we	at_the_time	gamble	a_great	took	Sa	Pinto
fitxatu_genuenean	*jokalaria*	*egoera*	*berezi*	*batean*	*aurkitzen_baitzen*	
we_hired	the_player	situation	special	in_a	was_because	

   "At the time we hired Sa Pinto we took a great gamble because [the player] was in a special situation."

2. **Extra mention (EM):** The system proposes a mention that is not present in the gold standard.

   × [Bestetik], Zestoako Udalak mozio bat onartu zuen aho batez AEKren kontrako prozesua salatzeko.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

√ Bestetik, Zestoako Udalak mozio bat onartu zuen aho batez AEKren kontrako prozesua salatzeko.

Bestetik Zestoako_Udalak mozio_bat onartu_zuen aho_batez
Moreover Zestoa_City_Council a_motion accepted unanimously
AEKren kontrako prozesua salatzeko
AEK against the_trial to_condem

"Moreover, Zestoa City Council accepted a motion unanimously to condemn the trial against AEK."

3. **Span errors (SE):** The system mention does not match exactly with the true or gold mention. Two different span error cases can happen.

3.1 **Head identified (HI):** The head of the mention to be identified is inside the span proposed by the system.

3.1.1 **Missing text (MT):** The system mention span misses one or more words that should be considered part of the gold mention.

× [Jiri Vanek] txekiarrak irabazi egin zion atzo.

√ [Jiri Vanek txekiarrak] irabazi egin zion atzo.

Jiri Vanek txekiarrak irabazi_egin_zion atzo
Jiri Van czech beat_him/her yesterday

"[Czech Jiri Van] beat him/her yesterday."

3.1.2 **Extra text (ET):** The system mention span has one or more words that are not part of the gold mention.

× [Luis Uranga harrituta] azaldu da Portugalgo klubaren jokaerarekin.

√ [Luis Uranga] harrituta azaldu da Portugalgo klubaren jokaerarekin.

"[Luis Uranga] is surprised with the attitude of the club from Portugal."

Luis Uranga harrituta_azaldu_da Portugalgo klubaren
Luis Uranga is_surprised from_Portugal of_the_club
jokaerarekin
with_the_attitude

"[Luis Uranga] is surprised with the attitude of the club from Portugal."

3.2 **Head not identified (HNI):** The head of the mention to be identified is not inside the system mention span.

× Europako Kontseiluetan herrialde handiek indar [gehiago] izango dute.

√ Europako Kontseiluetan herrialde handiek [indar gehiago] izango dute.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Europako Kontseiluetan herrialde handiek indar gehiago
European in_councils nations The_great strength more
izango_dute
will_have

"The great nations will have [more strength] in European councils."

4. **Divided mention (DM):** Mention to be identified is divided in two or more mentions.

× Eztandak ez du inor zauritu, eta kalte material txikiak eragin ditu, Hammersmith zubian, [Londres] [mendebaldean].
√ Eztandak ez du inor zauritu, eta kalte material txikiak eragin ditu, Hammersmith zubian, [Londres mendebaldean].

Eztandak ez_du inor zauritu eta kalte material txikiak
The_explosion have_not anyone hurt and damage material small
eragin_ditu Hammersmith_zubian Londres mendebaldean
has_been in_Hammersmith_Bridge London west

"The explosion in Hammersmith Bridge, [west London], has not hurt anyone and material damage has been small."

5. **Conflated mentions (CM):** Two or more mentions are identified as they were only one.

× Gogotsu ariko da [Euskaltel Alpeetako Klasikoan] eta Dauphine Liberen ariko dira hurrena.
√ Gogotsu ariko da [Euskaltel] [Alpeetako Klasikoan] eta Dauphine Liberen ariko dira hurrena.

Gogotsu ariko_da Euskaltel Alpeetako_Klasikoan
enthusiastically will_participate Euskaltel in_the_Classique_des_Alpes
eta Dauphine_Liberen ariko_dira hurrena
and in_Dauphine_Libere they_will_take_part then

"[Euskaltel] will participate enthusiastically in the Classique des Alpes and then they will take part in Dauphine Libere."

6. **Combination of errors (CN):** Two or more errors explained above can be identified in the mention proposed by the system. The following examples illustrate two cases of combined errors:

- **Divided Mention + Missing Text**

× [Siriako Atzerri ministro Farouk] Al [Xaraak] ukatu egin du Israelek Xebaako landetxeak okupatzea onartzen duela.
√ [Siriako Atzerri ministro Farouk Al Xaraak] ukatu egin du Israelek Xebaako landetxeak okupatzea onartzen duela.

Siriako Atzerri ministro Farouk Al Xaraak ukatu_egin_du
The_Syrian Foreign Minister Farouk al Sharaa has_denied
Israelek Xebaako landetxeak okupatzea onartzen duela
Israel of_Shebaa the_farms to_occupy accepts that

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 6. *Examples of error types*

Error type	System	Gold
Missing mention	–	-jokalaria
Extra mention	-Bestetik	–
Missing text	-Jiri Vanek	-Jiri Vanek txekiarrak
Extra text	-Luis Uranga harrituta	-Luis Uranga
Head not identified	-gehiago	-indar gehiago
Divided mention	-Londres	-Londres mendebaldean
	-mendebaldean	–
Conflated mentions	-Euskaltel Alpeetako Klasikoan	-Euskaltel
	–	-Alpeetako Klasikoan
Combination of errors	-Siriako Atzerri ministro Farouk -Xaraak	-Siriako Atzerri ministro Farouk Al Xaraak

"[The Syrian Foreign Minister Farouk al Sharaa] has denied that Israel accepts to occupy the farms of Shebaa."

- **Conflated Mentions + Extra text**

  × [Clintonen irudiko, Alderdi Errepublikarraren egitasmoa] ez da egokia europarren interesetarako.
  √ [Clintonen] irudiko, [Alderdi Errepublikarraren egitasmoa] ez da egokia europarren interesetarako.

  Clintonen irudiko Alderdi_Errepublikarraren egitasmoa ez_da
  Clinton believes of_Republican_Party the_project is_not
  egokia europarren interesetarako
  suitable of_Europeans for_the_interest

  "Clinton believes the project of the Republican Party is not suitable for the interest of Europeans."

Tables 6 summarises the examples explained above.

Table 7 shows the number of each error type when automatic mention detection is carried out and the percentages. The percentages are calculated taking into account the total number of correct mentions (1,904) in the subcorpus used for the evaluation.

As can be observed, the majority of errors are Missing Mentions (11.50%) and Extra Mentions (7.46%). The last row shows that the errors are 32.10% of the total correct mentions.

Each error type differently affects the precision and recall values. The differences also depend on the scoring protocol that has been used for scoring. For instance, the conflated mentions error type provokes one precision error and k (the number of mentions in which they have been conflated) recall errors when Strict Matching protocol is used. To better understand this, we can observe the following example, which has previously been presented.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 7. *Error types automatic mentions*

Error type	#	%
Missing mention	219	11.50
Extra mention	142	7.46
Missing text	103	5.41
Extra text	65	3.41
Head not identified	5	0.26
Divided mention	25	1.31
Conflated mentions	32	1.68
Combination of errors	20	1.05
Total	611	32.10

Table 8. *Effects of error types in P and R*

	Strict matching		Lenient matching	
Error type	P	R	P	R
Missing mention	0	1	0	1
Extra mention	1	0	1	0
Missing text	1	1	0	0
Extra text	1	1	1	1
Head not identified	1	1	1	1
Divided mention	k	1	k-1	0
Conflated mentions	1	k	1	k

× Frantziako ibilbide gogorretan gogotsu ariko da [Euskaltel Alpeetako Klasikoan] eta Dauphine Liberen ariko dira hurrena.
√ Frantziako ibilbide gogorretan gogotsu ariko da [Euskaltel] [Alpeetako Klasikoan] eta Dauphine Liberen ariko dira hurrena.

In this example, the system conflates two mentions in one ([Euskaltel Alpeetako Klasikoan]), so k= 2. This mention is considered incorrect when Strict Matching is applied as it does not exactly match with any mention in the gold, and also when Lenient Matching is applied. Although it has the head of the correct mention inside its span, it exceeds the span of the gold mention. Consequently, it is considered a precision error. Regarding recall, as k is two in this case, two recall errors are counted because [Euskaltel] and [Alpeetako Klasikoan] gold mentions are not present in the system output.

Table 8 summarises how error types affect precision and recall when different scoring protocols are used. In the case of combination of errors, the effect in precision and recall depends on the nature of the combined errors. The same happens when the scoring protocol is Lenient Matching.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

6.2 Error causes

To help identify the error causes, we carried out a qualitative evaluation on a corpus of 1,070 mentions, a subcorpus of the 1,904 mentions that were used to analyse the error types.

We divided the errors encountered in nine categories. For each category, we present some examples. The first item in each example is the incorrect case (×) and the second the respective correct one (√).

1. **Erroneous morphosyntactic analysis (EMA):** POS tags and/or syntactic functions are incorrectly disambiguated by the lemmatiser. The same string of letters can be a noun in some contexts and an adjective in others; this error occurs when a non-noun instance is tagged as a noun and, consequently, incorrectly regarded as a chunk. In addition, some complex postpositional are not identified correctly. In the following example, *arte* is tagged as a noun (*arte*, n.: art) and, consequently, considered a nominal chunk, although in this context it is a part of a complex postposition that has not been identified, meaning "until" which is a preposition in English.

   × Futbolean azken minutura [arte] edozer gauza gerta daiteke.
   √ Futbolean azken minutura arte edozer gauza gerta daiteke.

   Futbolean azken minutura arte edozer gauza gerta_daiteke
   In_football last the_minute until anything can_happen

   "In football, anything can happen until the last minute."

2. **Incorrect chunk tag (ICT):** The chunker does not provide correct chunk tags.

   × Estatu Batuetako presidente Bill [Clintonek] antimisilen inguruan Pentagonoa atontzen ari den sistema babesteko eskatu zion atzo Europar Batasunari Lisboan.
   √ [Estatu Batuetako presidente Bill Clintonek] antimisilen inguruan Pentagonoa atontzen ari den sistema babesteko eskatu zion atzo Europar Batasunari Lisboan.

   Estatu Batuetako presidente Bill Clintonek antimisilen_inguruan
   States United president Bill Clinton the_anti-missile
   Pentagonoa atontzen_ari_den sistema babesteko eskatu_zion
   the_Pentagon that_is_preparing system to_support asked
   atzo Europar Batasunari Lisboan
   Yesterday European the_Community in_Lisbon

   "Yesterday in Lisbon, [United States president Bill Clinton] asked the European Community to support the anti-missile system that the Pentagon is preparing."

3. **Missing chunk tag (MCT):** The chunker does not tag a chunk, or it only tags the beginning or the end of the chunk. Not opened or not closed tags can occur when CG formalism (Karlsson *et al.* 1995) is used.

   × [Luis Uranga harrituta azaldu da Portugalgo klubaren jokaerarekin.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

√ [Luis Uranga] harrituta azaldu da Portugalgo klubaren jokaerarekin.

Luis Uranga harrituta_azaldu_da Portugalgo klubaren
Luis Uranga is_susprised Portuguese of_the_club
jokaerarekin
at_the_attitude

"[Luis Uranga] is surprised at the attitude of the Portuguese club."

The error could be partially solved in the case when the chunker missed the ending tag of a chunk. In this case, we applied a heuristic that is explained in Section 6.4.

4. **Incorrect clause tag (ICLT):** The left clause boundary proposed by the clause tagger is not the same as the gold mention left boundary, so the mention is incorrect.

× Behin betiko hitzarmen bat sinatu behar dute bi aldeek, [horrela Oslon hasitako prozesua] bukatzeko.
√ Behin betiko hitzarmen bat sinatu behar dute bi aldeek, horrela [Oslon hasitako prozesua] bukatzeko.

Behin_betiko hitzarmen bat sinatu behar_dute bi aldeek horrela
permanent agreement a sign are_going_to two the_parts so_that
Oslon hasitako prozesua bukatzeko
in_Oslo that_began the_process will_be_brought_to_a_close

"The two parts are going to sign a permanent agreement so that [the process that began in Oslo] will be brought to a close."

5. **Predication (PRE):** Two words are joined in one chunk even though there is a predication relation between them. In coreference resolution it is common to join two mentions in one cluster when they have a predicational relation. For this reason, it is necessary to split the chunk into two parts when there is a predication. However, in some cases it is difficult to predict the predication relation and the chunk is not split, causing a mention detection error.

× [Bush buru] duen Alderdi Errepublikarraren proiektua okerragoa da.
√ [Bush] [buru] duen Alderdi Errepublikarraren proiektua okerragoa da.

Bush buru duen Alderdi Errepublikarraren proiektua
Bush leader whose_is Party of_the_Republican the_project
okerragoa da
worse is

"The project of the Republican party whose [leader] is [Bush] is worse".

The opposite case also occurs, i.e., a chunk is split when it should not be, causing a mention detection error.

× [Madariagaren] [ametsa] da aspalditik.
√ [Madariagaren ametsa] da aspalditik.

Madariagaren ametsa da aspalditik
Madariaga's dream it_has_been long

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

"It has long been [Madariaga's dream]."

The error is caused by a rule that does not divide a mention when it is presented in a predication relation or divides it when it should not. The rule can be adapted to better identify whether the mentions should be divided.

6. **Coordination (COR):** In coordination cases, the entire coordinate structure is considered a mention. However, in some cases the submentions to the left and right of the coordinating conjunction do not need to be marked. The detector should not split the coordinate structure when the submentions do not have to be included, otherwise we get incorrect mentions in the output.

   × [[Lintxamendu] eta [jazarpenerako erabili nahi den aitzakia]] da.
   √ [Lintxamendu eta jazarpenerako erabili nahi den aitzakia] da.

   Lintxamendu eta jazarpenerako erabili_nahi_den aitzakia da
   lynching and for_persecution to_use an_excuse is

   "It is [an excuse to use for lynching and persecution]."

   The same happens when the whole coordinate structure is not joined.

   × [Bost urteko kontratua] eta [sekulako dirutza] lortu zituen.
   √ [[Bost urteko kontratua] eta [sekulako dirutza]] lortu zituen.

   Bost_urteko kontratua eta sekulako dirutza lortu_zituen
   a_five-year contract and lots_of money He/she_obtained

   "He/she obtained [[a five-year contract] and [lots of money]]."

   In addition, sometimes coordinate structures are enumerations. In enumeration cases, apart from the entire coordinate mention, submentions divided by the comma must also be obtained as mentions. When these enumerations are not divided correctly mentions are missed.

   × [Van der Linden, Malafosse, Carboneau, Carrat, Venditti, Arbizu, Lamaison eta Smith] dira jokalariak.
   √ [[Van der Linden], [Malafosse], [Carboneau], [Carrat], [Venditti], [Arbizu], [Lamaison] eta [Smith]] dira jokalariak.

   Van der Linden Malafosse Carboneau Carrat Venditti Arbizu
   Van der Lindern Malafosse Carboneau Carrat Venditti Arbizu
   Lamaison eta Smith dira jokalariak
   Lamaison and Smith are the_players

   "The players are [[Van der Linden], [Malafosse], [Carboneau], [Carrat], [Venditti], [Arbizu], [Lamaison], and [Smith]]."

   Adapting the rules that decide whether a coordinate structure needs to be divided into submentions and joined into bigger structures would necessarily improve mention detection scores. In addition, specific rules to treat enumerations would also improve mention detection.

7. **Comparatives (COMP):** In comparatives structures, the related two elements are not joined and, consequently, the adverb (baino than) is considered out of the mention.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

× [Bilerak] baino [oihartzun handiagoa] lortu zuen atentatuak.
√[Bilerak baino oihartzun handiagoa] lortu zuen atentatuak.

Bilerak baino oihartzun handiagoa lortu_zuen atentatuak.
the_meeting than coverage greater received the_attack

"The attack received [greater coverage than the meeting]."

A specific rule in the mention detector to treat comparative structures would solve the problem.

8. **Referentiality (RF):** Some referential chunks are discarded as they do not have nouns inside them.

× Italiarrak onartezintzat hartu du erabakia.
√ [Italiarrak] onartezintzat hartu du erabakia.

Italiarrak onartezintzat hartu_du erabakia
The_Italian unacceptable considers the_decision

"[The Italian] considers the decision unacceptable."

On the contrary, non-referential chunks that have nouns inside are considered as mentions, for example, linking words.

× [Era berean], lurralde suntsitua berreraikitzeko nazioarteko laguntza eskatu zuen.
√ Era berean, lurralde suntsitua berreraikitzeko nazioarteko laguntza eskatu zuen.

Era_berean lurralde suntsitua berreraikitzeko
In_the_same_way territory the_destroyed to_rebuild
nazioarteko laguntza eskatu_zuen
for_international help he/she_called

"In the same way, he/she called for international help to rebuild the destroyed territory."

9. **Miscellaneous (MISC):** Errors with varied causes that cannot be classified into the categories discussed earlier. Some examples include:

- **Parenthetical:** The chunks are divided by a parenthetical structure.

× [David Trimble] ([UUP]) [lehen ministroak] adierazpenak egin zituen.

√ [David Trimble ([UUP]) lehen ministroak] adierazpenak egin zituen.

David Trimble (UUP) lehen ministroak adierazpenak egin_zituen
David Trimble (UUP) Prime Minister statements made

"[Prime Minister David Trimble ([UUP])] made statements."

- **Typing error:** Some texts have typing errors, missing spaces between words, unreadable characters... When processing these texts with automatic tools the typing errors are not always solved and cause a mention detection error.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

× [1995eanhartu] zuen parte.

√ [1995ean] hartu zuen parte.

[1995ean] hartu_zuen_parte
in_1995 he/she_participated

"He/she participated in [1995]."

Solving these problems is complicated, and they would be better corrected in the preprocessing step.

- **Extra Words in Boundaries:** The chunks contain words that fall outside the mention boundaries.

  × Kolpistek elkarrizketak hautsi zituzten [erabaki horren ondorioz].

  √ Kolpistek elkarrizketak hautsi zituzten [erabaki horren] ondorioz.

  Kolpistek *elkarrizketak* *hautsi_zituzten* *erabaki*
  The_leaders_of_the_coup the_talks broke_off decision
  horren_ondorioz
  because_of_that

  "The leaders of the coup broke off the talks because of that decision."

  The error of mentions with extra words can be solved by defining a set of words that appear at the boundaries of mentions even though they are not part of the mention. This way such words can be removed from within the mention boundaries.

- **Unjoined chunks:** Even though the chunks are correct they do not match the mentions exactly. In most cases, the chunks should be joined to create a correct mention.

  × Gentzelek [atzo] [goizean] hitz egin zuen.

  √ Gentzelek [atzo goizean] hitz egin zuen.

  Gentzelek *atzo* *goizean* *hitz_egin_zuen*
  Gentzel yesterday morning talked

  "Gentzel talked [yesterday morning]."

  The error of unjoined chunks that should form a mention can be solved by identifying the exact cases in which this phenomenon occurs.

Although specific rules to treat miscellaneous (MISC) cases would likely improve mention detection, the phenomenon is a broad one and the improvement may not be substantial.

We observed that the errors explained above can be classified in three main categories, some are caused by preprocessing tools (EMA, ICT, MST), some others are because of using the output of general-purpose tools that must be adapted to our mentions specification (ICLT, PRE, COR, CONP), while the rest are caused by inaccurate mention detection rules (RF). The MISC errors fall in all the categories.

Table 9 shows the percentages of the main error causes resulting from the mention detection step.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 9. *Main causes of errors*

	Error cause	#	%
1	EMA	47	14.11
2	ICT	93	27.93
3	MCT	63	18.92
4	ICLT	25	7.51
5	PRE	5	1.50
6	COR	35	10.51
7	COMP	2	0.60
8	RF	32	9.61
9	MISC	31	9.31
	Total	333	100.00

As Table 9 shows, the automatic preprocessing tools are the main source of errors: 27.93% are caused by the chunker providing incorrect chunk tags (ICTs) , and 18.92% by it missing the chunk tag. The morphosyntactic analysis that the mention detector receives as input also causes a number of errors in mention detection, accounting for 14.11% of the errors. Another considerable cause of error is that the left clause boundaries identified by the clause tagger do not match the left boundaries of gold mentions; this accounts for 7.51% of the errors.

Given the above, we can conclude that the input that our mention detector receives has a substantial impact on mention detection errors. This is also pointed out by Uryupina and Moschitti (2013), who suggest that better performance can be obtained if a robust preprocessing is achieved. The performance of the mention detector is highly dependent on the preprocessing tools, and errors caused by these tools are difficult to tackle. We can, however, attempt to address other types of errors.

We found that the greatest improvements can be achieved by treating the errors classified as COR and RF, which account for 10.51% and 9.61% of errors, respectively.

6.3 Distribution of errors among mention types

Apart from identifying the main causes of errors, observing how these errors are distributed among mention types is essential to implementing improvements effectively. Figure 4 shows this distribution in terms of percentages of the total number of errors.

As Figure 4 shows, ICT error cause is mainly divided into two mention types, proper nouns (10.81%) and common noun phrases (15.62%). A similar pattern can be found in the Missing Chunk Tag (MCT) errors. They cause 18.92% of the total errors and mainly affect common noun phrases (9.91%) and proper nouns (4.50%). Incorrect Clause Tag (ICLT) errors, in turn, are mainly errors in NPs containing subordinate clauses (6.91%). Other considerable errors are errors

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

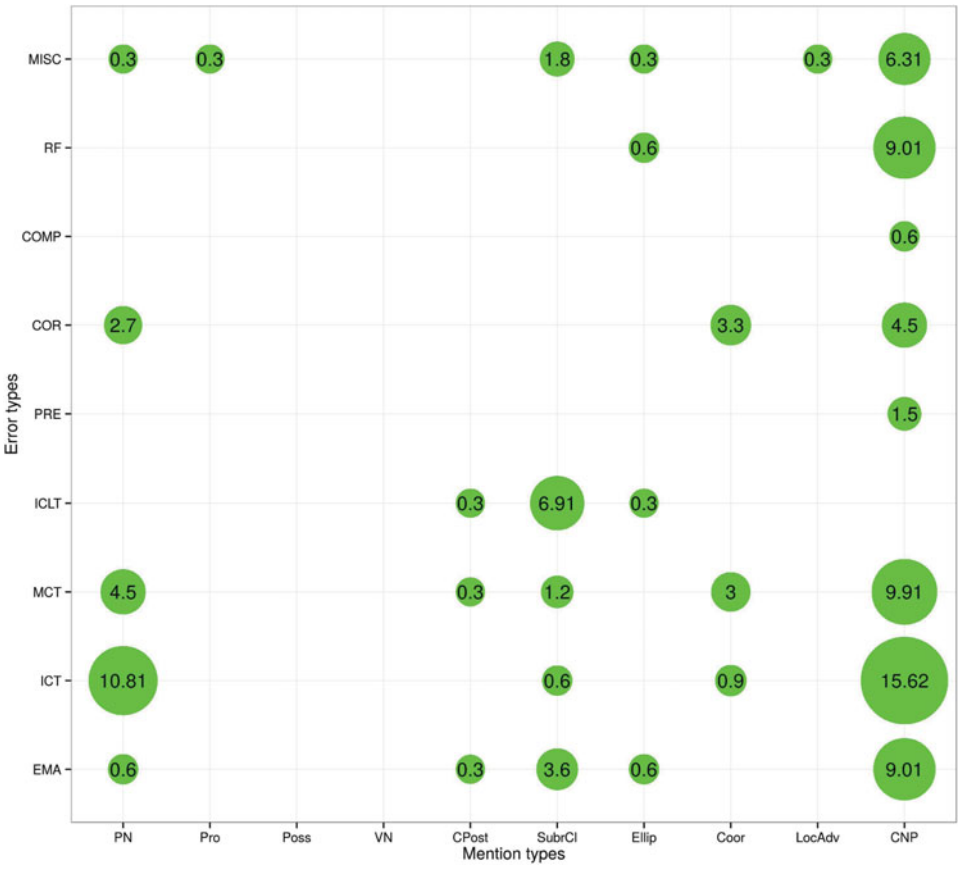


Fig. 4. (Colour online) Distribution of error causes into mention types. Automatic processing.

caused by Erroneous Morphosyntactic Analysis (EMA) (9.01%), Referentiality (RF) (9.01%), and MISC (6.31%) in Common Noun Phrases.

In the next Section, we explain how we improved our mention detection to overcome the problems identified by the error analysis.

6.4 Improvements in mention detection

To attempt to resolve some of the errors presented in Table 9, we implemented the following five adjustments and applied them in the mention detection process:

1. To solve the errors caused by the chunker only tagging the beginning of the chunk and not the end, we applied a heuristic that closes the nominal chunk when a verbal chunk starts. In other words, if a nominal chunk is not closed when a verbal one starts, we close it just before the verbal chunk. With this improvement, we tried to solve some of the *MCT* errors (third row in Table 9). In example (a), the nominal chunk *borroka armatua* "armed struggle" is opened properly but the close tag is missed. Since a verbal chunk *itzultzeko*

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

"to return" starts after the word *armatua* "armed", we close the nominal chunk, obtaining a correct nominal chunk as in example (b).

(a) IRA [borroka armatura itzultzeko asmotan da.

IRA borroka armatura itzultzeko asmotan_da
The_IRA struggle to_armed to_return intends

"The IRA intends to return to [armed struggle."

(b) IRA [borroka armatura] itzultzeko asmotan da.

IRA borroka armatura itzultzeko asmotan_da
The_IRA struggle to_armed to_return intends

"The IRA intends to return to [armed struggle]."

2. We observed in the error analysis that some mentions in a predication relation are divided when they should not be, or not divided when they should be. We adjusted the rule in the FSTs that was responsible for deciding whether a mention should be divided or not, depending on whether it was in a predication relation. With this improvement, we aimed to solve *PRE* errors (fifth row in Table 9). In the following examples, (a) is a case when a mention should be divided and (b) an example where the mention should not be divided.

(a) [Ibarretxe] [lehendakaria] da.

Ibarretxe lehendakaria da
Ibarretxe the_president is

"[Ibarretxe] is the [president]."

(b) [Bushen proposamena] da.

Bushen proposamen da
Bush's proposal it_is

"It is [Bush's proposal]."

3. We observed that chunks composed only of determiners were considered mentions. However, not all of these chunks are mentions. For instance, examples such as *biok* "we both" in (a) should be considered mentions, but examples such as *bestetik* "on the other hand" (b) should be discarded. We defined a new rule that discards chunks of undefined determiners because they are rarely referential. With this improvement, we aimed to detect some mentions that were missed and quantified as *RF* error (eighth row in Table 9).

(a) [Biok] egin genuen lan hura.

Biok egin_genuen lan_hura
We_both did the_work

"[We both] did the work."

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

(b) Bestetik, orain arte jasotako elkartasuna eskertu zuen.

Bestetik *orain_arte* *jasotako* *elkartasuna*
On_the_other_hand so_far he_had_received for_the_support
eskertu_zuen
he_expressed_his_thanks

"On the other hand, he expressed his thanks for the support he had received so far."

4. Chunks composed of only adjectives were all discarded. Upon closer examination, however, we observed that some of them could in fact be referential, as clarified by the examples below. In example (a), *handia* "big" is not referential so it must not be considered as a mention, but in example (b) the same word, *handia*, is referential and so is a mention. Therefore, we adjusted the rule that discarded chunks composed of nothing but adjectives so that it only discarded undefined adjectives and adjectives that were in a predication relation, as in example (a). With this improvement, our aim was to detect mentions that were missed and quantified as *RF* error (eighth row in Table 9).

(a) Etxea oso handia da.

Etxea *oso* *handia* *da*
The_house very big is

"The house is very big."

(b) [Handia] garestiagoa da.

Handia *garestiagoa* *da*
The_big_one more_expensive is

"[The big one] is more expensive."

5. Whether chunks to the left and the right of the coordinating conjunction are considered a single chunk or separate chunks is determined by the nature of the chunk, in particular whether the element(s) in the chunk are declined. In example (a), both Pello and Ane are declined (-ri to) and are therefore considered two separate chunks, but in (b) Pello is not declined (-Ø) and therefore the elements are considered a single chunk. We defined new rules that identify coordinate structures that are not joined but that should be considered mentions, e.g. (a). With this improvement, our goal was to detect coordinate structures that were missed, thereby decreasing *COR* error (sixth row in Table 9).

(a) [Pellori] eta [Aneri] oparia eman zioten.

Pellori *eta* *Aneri* *oparia* *eman_zioten*
to_Pello and to_Ane the_present they_gave

"They gave the present to [Pello] and [Ane]."

(b) [Pello eta Aneri] oparia eman zioten.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 10. *Improved scores in mention detection. BMD=Basque Mention Detector, BIMD=Basque Improved Mention Detector*

	Strict matching			Lenient matching		
	P	R	F_1	P	R	F_1
BMD	73.86	72.87	73.36	78.69	80.71	79.68
BIMD	74.67	74.47	74.57	79.26	81.92	80.57

Table 11. *Comparison of main error types when improvements in mention detection are applied*

	Automatic BMD		Automatic BIMD	
Error type	#	%	#	%
Missing mention	219	11.50	175	9.19
Extra mention	142	7.46	138	7.25
Missing text	103	5.41	102	5.36
Extra text	65	3.41	66	3.47
Head not identified	5	0.26	3	0.16
Divided mention	25	1.31	20	1.05
Conflated mentions	32	1.68	30	1.58
Combination of errors	20	1.05	14	0.74
Total	611	32.10	548	28.80

Pello eta Aneri oparia eman_zioten
Pello and to_Ane the_present they_gave

"They gave the present to [Pello] and [Ane]."

After implementing these five improvements, we obtained a gain of 1.21% in the F-measure using Strict Matching and of 0.89% with Lenient Matching. These gains are significant.[3] The results can be seen in Table 10, and show that the improvements have succeeded in resolving some errors.

Table 11 shows the main error types with improved mention detection (BIMD) and without it (BMD). In absolute values, 61 errors have been resolved, 611 without improvements and 548 after applying them. The biggest improvement has been achieved in Missing Mention error type, 44 mentions that before were not identified are now in the system output and are correct mentions. In percentage values, an improvement of 3.3% is obtained, 28.80% being errors of total mentions.

7 Experiment with gold input

Given that the input received by the mention detector is a significant source of errors, we decided to experiment with what would happen with a perfect input. In

[3] Statistical significance is tested with Paired Student's t-test. p value < 0.01

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 12. *Mention detection scores with partially gold annotation*

	Strict matching			Lenient matching		
	P	R	F_1	P	R	F_1
$BIMD_a$	74.67	74.47	74.57	79.26	81.92	80.57
$BIMD_g$	84.95	86.84	85.89	87.06	91.15	89.06

Table 13. *Comparison of main error types when partially gold input is provided*

	Automatic BMD		Automatic BIMD		Gold BIMD	
Error type	#	%	#	%	#	%
Missing mention	219	11.50	175	9.19	88	4.62
Extra mention	142	7.46	138	7.25	87	4.57
Missing text	103	5.41	102	5.36	33	1.73
Extra text	65	3.41	66	3.47	57	3.00
Head not identified	5	0.26	3	0.16	0	0.00
Divided mention	25	1.31	20	1.05	14	0.74
Conflated mentions	32	1.68	30	1.58	2	0.11
Combination of errors	20	1.05	14	0.74	4	0.21
Total	611	32.10	548	28.80	285	15.00

"real-world" cases, the mention detector is integrated into a pipeline in which errors made in each step affect the following phases. Providing a partially gold input to our mention detector enables us to evaluate the accuracy of our system in an isolated way. Due to the lack of a corpus where all linguistic levels are manually tagged, we experimented with gold lemmas, gold PoS tags, and gold chunks. The syntactic function, the named entity tags and clause tags are automatic, so the input is only partially gold. Table 12 shows the results obtained using partially gold information as input.

Using gold lemmas, PoS tags, and chunks, the scores clearly increase—in fact, they are highly promising. The F-measure value is 85.89% with Strict Matching and 89.06% with Lenient Matching. Comparing this to the results obtained by applying mention detection to automatic annotation, as shown in the first row, the results increase by 11.42% percentage points using Strict Matching and by 8.40% with Lenient Matching.

In view of these results, it is clear that the type of input provided is crucial. As Table 13 shows in last row, when partially gold input is used instead of automatic preprocessed input, the error types drop from 548 (Automatic BIMD) to 285 (Gold BIMD). This is a significant drop and clearly shows that the output of preprocessing tools has a substantial effect on mention detection. In percentages, the difference is 13.80%, so about 14% of errors are produced by preprocessing tools.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Table 14. *F1 scores of coreference resolution with different metrics*

	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL
B	25.17	54.39	51.62	51.92	28.13	43.83
BMD	28.36	61.62	57.96	58.91	35.56	49.63
BIMD	29.28	62.65	58.69	59.65	36.77	50.53

8 Effects of mention detection results in coreference resolution

In order to evaluate, how the improvements we obtained in mention detection affect coreference resolution, we used BART, a highly modular toolkit for developing coreference applications (Versley *et al.* 2008).

In our evaluation experiments, a simple coreference model has been used, namely, the mention-pair approach presented in Soon, Ng and Lim (2001). The metrics used to evaluate the system performance are MUC (Vilain *et al.* 1995), B^3 (Bagga and Baldwin 1998), $CEAF_e$ (Luo 2005), $CEAF_m$ (Luo 2005), and BLANC (Recasens and Hovy 2011). The CoNLL metric 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).

Table 14 shows the results obtained in coreference resolution when different mention detection systems are used , Baseline (B), Basque Mention Detector (BMD), and Basque Improved Mention Detector (BIMD).

BMD outperfoms the Baseline (B) scores in all the metrics. In CoNLL metric, BMD obtains a score of 49.63, which is 5.8 points higher than Baseline, which scores 43.83. The same happens when we compare BMD results with those obtained by BIMD. BIMD outperforms in all the metrics and CoNLL score is outperformed by 0.9. All the improvements obtained when different mention detectors are used are significant.[4]

9 Conclusions

We have presented a mention detector system for Basque. The system has been created based on a linguistic analysis of mentions in Basque. We explain these structures and analyse their presence in the EPEC corpus; we also compare them with other languages.

Good results at the mention detection stage have a major impact on the success of coreference resolution systems, and therefore improving mention detection is crucial. We applied several different strategies in an attempt to obtain more accurate mentions.

First, we concluded that using only generic preprocessing tools for mention detection does not produce good results (F-measure of 64.28% using Strict Matching and 71.12% using Lenient Matching). As a consequence, we created several FSTs

[4] Statistical significance is tested with Paired Student's t-test. p value < 0.01

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

that make use of generic tools. These FSTs obtain better results: F-measure of 73.36% using Strict Matching and 79.68% when Lenient Matching is used. In other words, this approach outperforms the baseline by nine points for Strict Matching and by eight points for Lenient Matching.

Secondly, we carried out a deep error analysis to better understand and thereby perhaps mitigate the main causes of errors in mention detection. This error analysis enabled us to define five improvements, which increased the F-measure by 1.21 points using Strict Matching (74.57%), and by 0.89 points using Lenient Matching (80.57%).

We also considered it important to quantify the impact that automatic preprocessing tools have on mention detection. Thus, we carried out an experiment with partially gold input. The results revealed that errors decrease substantially when gold input is provided: the better the preprocessing tools, the better the mention detection scores. The F-measure value obtained with gold input ($BIMD_g$) is 85.89% with Strict Matching and 89.06% with Lenient Matching—a difference of 11.42 points for Strict Matching and of 8.48 points for Lenient Matching in comparison with results obtained when automatic input is used ($BIMD_a$).

We integrated the mention detector system into an end-to-end coreference resolution system for written Basque and analysed the influence that mention detection has in coreference resolution results.

References

Aduriz, I., Aranzabe, M., Arriola, J. M., Atutxa, M., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and Urizar, R. 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. *Language and Computers* **56**(1): 1–15.

Aduriz, I., and Díaz de Ilarraza, A. 2003. Morphosyntactic disambiguation and shallow parsing in computational processing of Basque. In B. Oyharçabal (ed.), *Inquiries into the lexicon-syntax relations in Basque*, pp. 1–21. University of the Basque Country, Bilbao, Spain.

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. 2004. Representation and treatment of multiword expressions in Basque. In *ACL Workshop on Multiword Expressions (MWE '04)*, Barcelona, Spain, pp. 48–55.

Alegria, I., Aranzabe, M., Ezeiza, N., Ezeiza, A., and Urizar, R. 2002. Using finite state technology in natural language processing of Basque. In B. W. Watson, and D. Wood *Implementation and Application of Automata*, pp. 1–12. Lecture Notes in Computer Science. Berlin: Springer/Heidelberg.

Alegria, I., Artola, X., Sarasola, K., and Urkia, M. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing* **11**(4): 193–203.

Alegria, I., Ezeiza, N., Fernandez, I., and Urizar, R. 2003. Named entity recognition and classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información (JOTRI 2003)*, Madrid, Spain, pp. 198–203.

Arrieta, B. 2010. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean.* PhD Thesis, Computer Languages and Systems, University of the Basque Country, Donostia-San Sebastián, Spain.

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* **34**(4): 555–96.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Bagga, B., and Badlwin, B. 1998. Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, Granada, Spain, pp. 563–66.

Broscheit, S., Poesio, M., Ponzetto, S. P., Rodriguez, K. J., Romano, L., Uryupina, O., Versley, Y., and Zanoli, R. 2010. BART: a multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden, pp. 104–07.

Chang, K.-W., Samdani, R., Rozovskaya, A., Rizzolo, N., Sammons, M., and Roth, D. 2011. Inference protocols for coreference resolution. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task (CoNLL 2011)*, Portland, Oregon, pp. 40–4.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. 2004. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of Language Resources and Evaluation Conference (LREC 2004)*, Lisbon, Portugal, pp. 837–40.

Hacioglu, K., Douglas, B., and Chen, Y. 2005. Detection of entity mentions occuring in english and Chinese text. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, Vancouver, British Columbia, Canada, pp. 379–86.

Hulden, M. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, Athens, Greece, pp. 29–32.

Karlsson, F., Voutilainen, J., Heikkilä, J., and Anttila, A. 1995. *Constraint Grammar: Language-independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter.

Klein, D., and Manning, C. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL '03)*, Sapporo, Japan, pp. 423–30.

Kopeć, M., and Ogrodniczuk, M. 2014. Inter-annotator agreement in coreference annotation of Polish. In J. Sobecki, V. Boonjing, and S. Chittayasothorn, (eds.), *Advanced Approaches to Intelligent Information and Database Systems*, Studies in Computational Intelligence, vol. 551. Switzerland: Springer. Springer International Publishing, Switzerland.

Kummerfeld, J. K., Bansal, M., Burkett, D., and Klein, D. 2011. Mention detection: heuristics for the OntoNotes annotations. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task (CoNLL 2011)*, Portland, Oregon, pp. 102–6.

Kummerfeld, J. K., and Klein, D. 2011. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, Washington, USA, pp. 265–77.

Laka, I. 1996. A brief grammar of Euskara, the Basque language. http://www.ehu.es/grammar. University of the Basque Country.

Lalitha, D. S., Sundar, R. V., and Rao, R. K. P. 2014. A generic anaphora resolution engine for Indian languages. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, pp. 1824–33.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. 2013. A generic anaphora resolution engine for Indian languages. *Computational Linguistics* **39**(4): 885–916.

Luo, X. 2005. On coreference resolution performance metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, Vancouver, British Columbia, Canada, pp. 25–32.

Marcus, M., Marcinkiewicz, M., and Santorini, B. 1993. Building a large annotated corpus of english: the Penn treebank. *Computational Linguistics* **19**(2): 313–30.

Màrquez, L., Recasens, M., and Sapena, E. 2013. Coreference resolution: an empirical study based on SemEval-2010 shared task 1. *Language Resources and Evaluation* **47**(3): 661–94.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Mihăltz, M. 2008. Knowledge-based coreference resolution for Hungarian. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

MUC-6 1995. Coreference Task Definition (v2.3, 8 Sep 95). In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, Columbia, Maryland, USA, pp. 335–44.

MUC-7 1998. Coreference task definition (v3.0, 13 Jul 97). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, USA.

NIST 2008. Automatic Content Extraction 2008 Evaluation Plan (ACE08).

Nguy, G., Novák, V., and Żabokrtský, Z. 2009. Comparison of classification and ranking approaches to pronominal anaphora resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, London, UK, pp. 276–85.

Nguyen, N., Kim, J.-D., and Tsujii, J. 2008. Challenges in pronoun resolution system for biomedical text. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, pp. 2408–12.

Ogrodniczuk, M., and Kopeć, M. 2011. End-to-end coreference resolution baseline System for Polish. In Z. Vetulani (ed.), *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Oznań, Poland, pp. 167–71.

Ohta, T., Pyysalo, S., Tsujii, J., and Ananiadou, S. 2008. Open-domain anatomical entity mention detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse (ACL'12)*, Jeju, Republic of Korea, pp. 27–36.

Orasan, C., Cristea, D., Mitkov, R., and Branco, A. 2008. Anaphora resolution exercise: an overview. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Pradhan, S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. 2007. OntoNotes: a unified relational semantic representation. In *Proceedings of the International Conference on Semantic Computing (ICSC 2007)*, Irvine, California, pp. 517–26.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. 2012. CoNLL-2012 Shared task: modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the 16th Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea, pp. 1–40.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. 2011. CoNLL-2011 shared task: modeling unrestricted coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task (CoNLL 2011)*, Portland, Oregon, pp. 1–27.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, V. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, Maryland, pp. 30–5.

Recasens, M., and Hovy, E. 2010. Coreference resolution across Corpora: languages, coding Schemes, and preprocessing information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, Uppsala, Sweden, pp. 1423–32.

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

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. 2010. SemEval-2010 task 1: coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden, pp. 1–8.

Recasens, M., and Martí, M. 2010. AnCora-CO: coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation* **44**(4): 315–45.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* **27**(4): 521–44.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press

Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., and Díaz de Ilarraza, A. 2012. Mention detection: first steps in the development of a Basque correference resolution system. In *KONVENS 2012, The 11th Conference on Natural Language Processing*, Vienna, Austria, pp. 128–36.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. 2009. Conundrums in noun phrase coreference resolution: making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore, pp. 656–64.

Uryupina, O. 2008. Error analysis for learning-based coreference resolution. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Uryupina, O. 2010. Corry: a system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden, pp. 100–3.

Uryupina, O., and Moschitti, A. 2013. Multilingual mention detection for coreference resolution. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 100–8.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding (MUC6)*, Columbia, Maryland, pp. 45–52.

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. 2008. BART: a modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, Columbus, Ohio, pp. 9–12.

Zhekova, D., and Kübler, S. 2010. UBIU: a language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden, pp. 96–9.

https://doi.org/10.1017/S1351324916000206 Published online by Cambridge University Press