Coreference Resolution for the Basque Language with BART

ANDER SORALUZE, OLATZ ARREGI, XABIER ARREGI and ARANTZA DÍAZ DE ILARRAZA
UNIVERSITY OF THE BASQUE COUNTRY

MIJAIL KABADJOV and MASSIMO POESIO
UNIVERSITY OF ESSEX

ABSTRACT

We present our work on Coreference Resolution in Basque, a unique language which poses interesting challenges for the problem of coreference. We explain how we extend the coreference resolution toolkit, BART, in order to enable it to process Basque. Then we run four different experiments showing both a significant improvement by extending a baseline feature set and the effect of calculating performance of hand-parsed mentions vs. automatically parsed mentions. Finally, we discuss some key characteristics of Basque which make it particularly challenging for coreference and draw a road map for future work.

INTRODUCTION

- Basque differs considerably in grammar for the languages spoken in surrounding regions
- It is agglutinative, head-final, prod-drop, free-word order language
- It has genderless and anymacyless system for pronouns
- Preliminary work on Coreference for Basque (Soraluze et al., 2015)

ANNOTATED CORPUS OF BASQUE

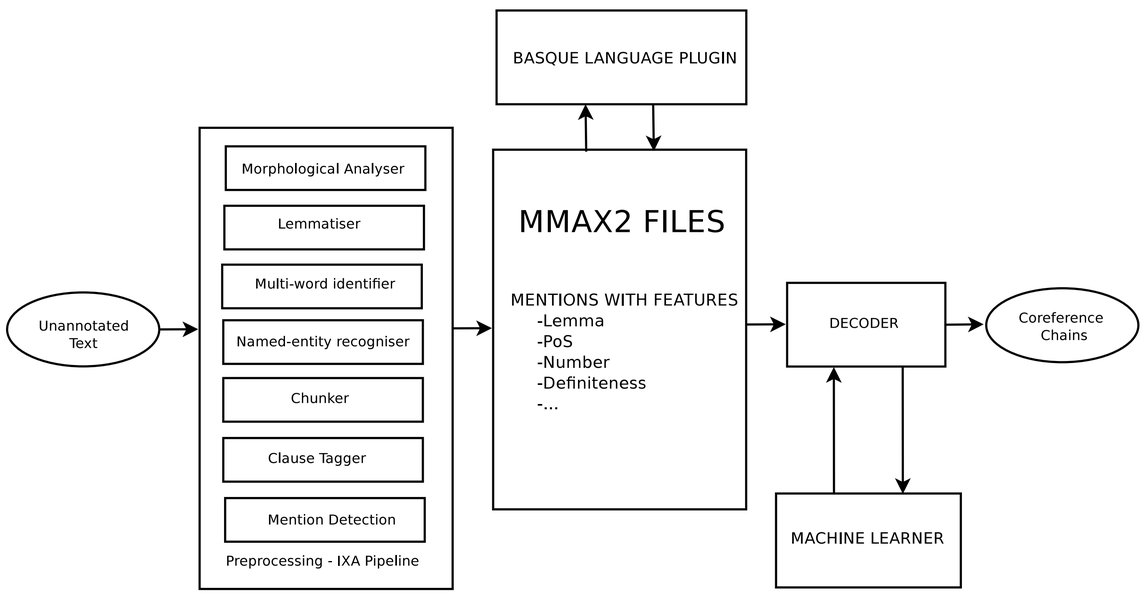
- EPEC, the reference corpus for the processing of Basque
- 300,000 word sample collection
- Manually annotated at different levels
- Mention and coreference chains annotated
- EPEC-coref corpus information

	Words	Mentions	Clusters	Singletons
Train	23520	6525	1011	3401
Devel	6914	1907	302	982
Test	15949	4360	621	2445

EXTENDING BART TO BASQUE

- BART's flexible modular architecture ensures its portability to other languages
- An independent *Language Plugin* module handles language specific information
- Adaptation process:
  - Used a preprocessing pipeline of Basque linguistic processors
  - Developed the *Basque Language Plugin*

SYSTEM ARCHITECTURE



MODELS

- Trained BART with two different models
- Baseline (Soon et al., 2001) model and model with Basque oriented features
- Experimentation with automatic and gold mentions

Features		Baseline	Basque
Gender	M_i and M_j agree in gender	$\surd$	$\surd$
Number	M_i and M_j agree in number	$\surd$	$\surd$
Alias	Matches abbreviations and name variations	$\surd$	$\surd$
StringMatch	M_i and M_j have the same surface form	$\surd$	$\surd$
SemClassAgree	Assesses the semantic compatibility of M_i and M_j	$\surd$	$\surd$
Appositive	M_i and M_j are in apposition structure	$\surd$	$\surd$
DistanceSentence	Distance in sentences between M_i and M_j	$\surd$	$\surd$
LemmaMatch	M_i and M_j have the same surface lemma	$\times$	$\surd$
HeadMatch	M_i and M_j have the same head	$\times$	$\surd$
StringKernel	Computes the similarity M_i and M_j strings	$\times$	$\surd$
DistanceMarkable	Distance in markables between M_i and M_j	$\times$	$\surd$
HeadPartofSpeech	M_i and M_j head PoS are the same	$\times$	$\surd$

EXPERIMENTAL RESULTS

		AUTOMATIC			GOLD		
		R	P	F_1	R	P	F_1
Mention Detection		72.91	74.69	73.79	100	100	100
MUC	Soon	18.37	67.23	28.86	23.62	78.66	36.34
	Basque	35.44	45.53	39.86	49.49	57.28	53.10
B^3	Soon	53.96	72.85	62.00	74.66	98.00	84.75
	Basque	58.10	65.27	61.48	81.21	87.78	84.37
$CEAF_m$	Soon	57.50	58.90	58.19	75.58	75.58	75.58
	Basque	58.67	60.10	59.38	76.59	76.59	76.59
$CEAF_e$	Soon	67.42	52.93	59.31	91.11	70.29	79.35
	Basque	61.63	58.15	59.84	82.10	77.64	79.81
BLANC	Soon	32.29	62.47	36.46	57.08	89.79	61.68
	Basque	38.70	48.81	42.41	66.78	75.99	70.34
CONLL	Soon	-	-	50.05	-	-	66.81
	Basque			53.72			72.42

ERROR ANALYSIS

Mention Detection error

- **Gold mentions**: *[Del Bosquek] prentsaurrekoa eman zuen atzo. [Vicente Del Bosque], [Real Madrileko entrenatzailea], nahikoa kezkati azaldu zen.*
- **Automatic mentions**: *[Del Bosquek] prentsaurrekoa eman zuen atzo. [Vicente Del Bosque , Real Madrileko entrenatzailea], nahikoa kezkati azaldu zen.*
- [Del Bosque] gave a press conference yesterday. [Vicente Del Bosque], [Real Madrid coach], appeared quite concerned.

Missed or incorrectly resolved pronoun

- *Lehendakari hautatu zutenetik, [Djukanovicek] aldaketa handia eman dio [bere] ildo politikoari.*
- Since he was elected as president, [Djukanovic] has greatly changed [his] policy lines.

Challenging case of pronoun

- *Gobernuaren bilera honen ondoren, oporretara joango da [Jospin], eta hauek baliatuko ditu, ziur aski, Chevenement kasuaz gogoetak egiteko eta konponbide batekin [bere] jarduerari eusteko.*
- After this government meeting, [Jospin] will go on holidays, and will surely use it to reflect on Chevenement case and to maintain [his] activity with a new solution.

Correctly resolved pronoun

- *"[Guk] ez dugu inoiz penaltietan irabazi." Luzapena golik gabe amaitzean, itzal beltz batek estali zuen Arena estadioa. Rijkaard-ek esana zuen arreta bereziz prestatu zituztela penaltiak, "[gure] istoria ez errepikatzeko."*
- "[We] have never won on penalties." After the extension finished without goals, a large shadow turn off the stadium. Rijkaard said they prepared penalties with great attention,"so that [our] story would not occur again".

CONCLUSIONS AND FUTURE WORK

Conclusions:

- Presented ongoing work or Coreference Resolution in Basque
- Adapted BART to enable it to process Basque
- Ran two levels of experiments: automatic vs. gold mentions
- Two different models, baseline model and Basque model
- Basque model significantly outperforms the baseline
- Discussed key characteristics for Basque Coreference Resolution

Future work:

- Investigate features that can make up for the lack of gender and animacy
- Extrinsic evaluation guaging the effect of coreference on a higher level task

- Ander Soraluze, Olatz Arregi, Xabier Arregi, 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.
- Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4):521–544, December.