Contribution of complex lexical information to solve syntactic ambiguity in Basque

Aitziber Atuxa Eneko Agirre Kepa Sarasola



lxa Group, University of the Basque Country (UPV/EHU) Manuel Lardizabal pasealekua, $1 \cdot 20018$ Donostia-San Sebastián

{kepa.sarasola@ehu.es}

11th December 2012

Syntactic ambiguity: PP attachment

- PP attachment is one of the most frequent syntactic ambiguities in English.
- Example:
 - "I saw the man with the telescope"
 - 2 different interpretations:
 - 1. I saw [the man] [with the telescope]
 - 2. I saw [the man [with the telescope]]

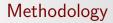
Syntactic ambiguity. Motivation

Syntactic ambiguities differ from language to language

Ambiguity	English	Basque
PP-attachment	50%	0.01%
Subj-Obj	-	33%

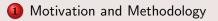
Example of subject/object ambiguity in Basque:

 "Bekak jaso ditu" Bekak jaso ditu. grant-abs-pl/erg-sg? get trans-aux+agr(he,them)
 2 different interpretations: "The grant-subj got (them)." "(He) got the grants-obj."



Our goal: Parse correction

- Focus on solving a relevant ambiguity
- Build a classifier using some features to solve it
- Replacing parser's result on ambiguous relations by the results of the classifier



- 2 Subject-object ambiguity in Basque
- 3 Features involved in the subject-object ambiguity
- 4 Experimental setup and Evaluation

6 Related work



6 Conclusions and future work

Subject Object ambiguity in Basque

- Morphologically rich, free word order languages (MoR-FWO):
 Czech, Turkish, Hindi...
- MoR-FWO Ergative Languages.
 - 2 different cases for marking subjects: Absolutive and Ergative.
 - Basque, Hindi and Urdu, Georgian, Tibetan, Eskimo...
- In Basque:

$$absolutive = \left\{ egin{array}{ccc} subject & of & intransitive & verbs \ object & of & transitive & verbs \end{array}
ight.$$

Subject Object ambiguity in Basque (Examples)

- Finite sentences: auxiliary marks transitivity
 - Bere beka bukatu da. His grant-abs-sg end intrans-aux+agreement(it).
 "His grant-subj has ended."
 - 2. Ø beka jaso zuen. ellided pro grant-abs-sg get trans-aux+agreement(he,it). "(He) got a grant-obj."
- But the ambiguous suffix -ak can mean absolutive plural or ergative singular.

3. Ø bekak jaso ditu. ellided pro grant-abs-pl/erg-sg? get trans-aux+agr(he,them). "The grant-subj got (them)." ?? "(He) got the grants-obj." ??

Subject Object ambiguity in Basque (Examples)

Examples

- In infinite sentences (lack of auxiliary marking transitivity) absolutive elements are ambiguous between subject and object
- 4. [Krisia bukatzea] espero dugu. Crisis-abs-sg finish-to hope transitive-aux-we.

"We hope that the crisis-subj will finish."

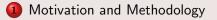
 \ldots (but in the Basque sentence "will finish" is an infinitive form)

5. Ø [Krisia gainditzea] espero dugu. ellided pros crisis-abs-sg overcome-to hope transitive-aux-we. "(We) hope (anyone/we) to overcome the crisis-obj ".

Subject Object ambiguity in Basque

- Depending on the transitivity of the verb the absolutive case will be subject or object.
- But the transitivity of the verb changes depending in the context.
 - Many verbs show transitivity alternations. For example "to break":
 - "I broke the window."
 - "The window broke."
 - The transitivity of certain verbs depends on their meaning For example as in English "to leave ":
 - intrans: "The train leaves at 5 o'clock. "
 - trans: "The hurricane left a trail of devastation."

Acquisition of verbal subcategorization information



- 2 Subject-object ambiguity in Basque
- B Features involved in the subject-object ambiguity
- 4 Experimental setup and Evaluation

6 Related work



6 Conclusions and future work

Acquisition of verbal subcategorization information

Features involved in the subject-object ambiguity

Subject-object ambiguity is associated to linguistically well motivated features

- Features related to morphological and syntactic information in the sentence
 - Preverbal position?
 - Ergative case?
 - infinitive?
 - ...
- Features related to verbal subcategorization information on the main verb(transitivity)

Acquisition of verbal subcategorization information

Feature Space

Features related to other morphological and syntactic information in the sentence

- AspectCtrl: 1 if the governing verb is a control/aspect verb (begin, stop, end, want, etc)
 - I started [PRO knowing you]. Infinitival without subject
- Preverb: 1 if the ambiguous element is in the preverbal position
- Inf: 1 if the verb appears in infinitival form
- *Erg*: 1 if the case is ergative
- -ak: 1 if the element bears the ambiguous -ak morpheme
- Sing: 1 if the element shows up in singular form
- *Entity*: 1 if the element is an entity

Acquisition of verbal subcategorization information

Features involved in the subject-object ambiguity

Acquisition of verbal subcategorization information

4 main sources

- Subcategorization Dictionary obtained from monolingual Basque corpus
- Queries over the Web
- Queries over an English parsed corpus
- Traditional Basque dictionary

Acquisition of verbal subcategorization information

Features involved in the subject-object ambiguity

Source: Subcategorization Dictionary

- Automatically built from raw corpora (10M words)
- Using a chunker + small grammar (78% phrases were correctly attached to verbs)
- We collected the following frequencies for each verb:
 - overall transitivity
 - noun-case-verb triplets
 - noun-case-verb-transitivity tuples

Acquisition of verbal subcategorization information

Features involved in the subject-object ambiguity

Source: Web as a corpus

For each Basque ambiguous noun-verb candidate:

- Construct all possible element+case+verb+auxiliary tuplets (aprox. 120)
 - Generate all possible subject-object unambiguous inflected forms (element+case)
 - Generate the 3 different inflected forms of the main verb
 - Generate the corresponding transitive-intransitive auxiliary forms (20 most frequent)
- Search in Google and get hits

Acquisition of verbal subcategorization information

Features involved in the subject-object ambiguity

Source: English monolingual corpus

BNC corpus parsed (10M verb-noun relations using RASP parser) **Assumption**: subject-object relation is stable across languages

For each Basque ambiguous noun-verb candidate:

- Translate the dependent lemma and the verb lemma using a bilingual dictionary
- Build all possible translation pairs
- Collect hits of each pair as subject and as object in the English corpus

Acquisition of verbal subcategorization information

Features involved in the subject-object ambiguity

Source: Traditional Basque dictionary

each verbal entry encodes the transitivity for each sense

We just considered the first sense

7 different markers for transitivity and transitivity alternations:

da, zaio, da/zaio, du, du/dio, dio, du/da.

- da, zaio, da/zaio: intransitive
- du, du/dio, dio: transitive
- du/da: transitive (intransitive with inchoative alternation)

Acquisition of verbal subcategorization information

Feature Space

8 features related to subcategorization

TransCase(SubcatDict)

The probability of the element to be a subject based on:

case: actual case assigned by the morphological analyzer
 P(TransCase): probability of the verb to be transitive according to the subcategorization dictionary

$$\label{eq:asympt} \textit{TransCase}(\textit{SubcatDict}) \left\{ \begin{array}{ll} \textit{P}(\textit{TransCase}) = \frac{\#\textit{trans}}{\#\textit{trans} + \#\textit{intrans}} & \textit{case} = \textit{erg} \& \textit{P}(\textit{TransCase}) > 0.5 \\ 1 - \textit{P}(\textit{TransCase}) & \textit{case} = \textit{abs} \& \textit{P}(\textit{TransCase}) < 0.5 \\ 0 & \textit{case} = \textit{abs} \& \textit{P}(\textit{TransCase}) > 0.5 \\ none & \textit{otherwise} \end{array} \right.$$

 TransCase(Web) equivalent to TransCase(SubcatDict) but based on the web frequencies

Acquisition of verbal subcategorization information

Feature Space

Features related to subcategorization

NCaseV(SubcatDict)

The probability of the element to be a subject based on:

- **case**: probability of that element to bear ergative with that verb
- P(TransCase): probability of the verb to be transitive according to the subcategorization dictionary

```
NCaseV(SubcatDict) \left\{ \begin{array}{ll} 1 & P(TransCase) > 0.5 \& P(Erg) > 0.5 \\ 0 & P(TransCase) < 0.5 \& P(Erg) < 0.5 \\ none & otherwise \end{array} \right.
```

 NCaseV(Web) equivalent to NCaseV(SubcatDict) but based on the web frequencies

Acquisition of verbal subcategorization information

Feature Space

Features related to subcategorization

NCaseVAux(SubcatDict)

The probability of the element to be a subject based on probability of that element:

- to bear ergative with that verb and a transitive auxiliary
- to bear absolutive case with that verb and an intransitive auxiliary

$$NCaseVAux(SubcatDict) \left\{ \begin{array}{c} \frac{\#(n+abs+v+intransAux)+\#(n+erg+v+transAux)}{\#(n+case+v)} & \#(n+case+v) > 0\\ none & otherwise \end{array} \right.$$

 NCaseVAux(Web) equivalent to NCaseVAux(SubcatDict) but based on the web frequencies

Acquisition of verbal subcategorization information

Feature Space

Features related to subcategorization

Subj(BNC)

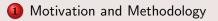
Value based on the probability of element's translation to be subject of verb's translation in BNC corpus:

 $Subj(BNC) \left\{ \begin{array}{ll} 1 & Prob(elementTranslation + subj) > 0 \\ 0 & Prob(elementTranslation + obj) > 0 \\ none & \notin BNC \land \neg translation \end{array} \right.$

TransCase(Dict)

Value based on the **actual case** and **transitivity** of the verb according to the Basque Monolingual Dictionary.

 $TransCase(Dict) \begin{cases} 1 & erg + (du || du - dio) \land abs + (da || da - zaio || zaio) \\ 0 & abs + (du || du - dio) \\ none & otherwise(du - da) \end{cases}$



- 2 Subject-object ambiguity in Basque
- B Features involved in the subject-object ambiguity
- 4 Experimental setup and Evaluation

6 Related work



6 Conclusions and future work

Experimental setup

Creating the gold standard

The gold Standard comprises 4,525 instances of ambiguous dependents in 3,617 sentences from around 11,000 sentences in the whole treebank

Steps to identify ambiguous elements:

- Ist look up the verbs. Depending on the finiteness there are two cases
 - finite forms: verb + auxiliary. Auxiliary resolves ambiguities except -ak cases
 if the subject and the object bear -ak auxiliary does not disambiguate

infinite form: dependents bearing absolutive are ambiguous

identify dependents and their cases to apply the previous rules

・ロッ ・行 ・ ・ ヨッ・ ・ ヨッ

Experimental setup

Methods

- The learning process:
 - Using the features described before we built a SVM classifier
 - The 4,525 relations in the Gold were divided in 2 sets: training (50%) and test (50%)
- The development over the train set
 - We evaluated each feature on its own
 - We evaluated the SVM classifier (cross-validation)
 - We performed feature ablation: learning with all features but one/some
- The evaluation against MaltParser (Final evaluation)
 - We compared our system with MaltParser over the test set

Evaluation on TRAIN

Results

- Baseline: assigning always the object tag, since it is the most frequent tag (75% Obj; 25% Subj)
- Evaluation of each feature on its own¹:

Feature	асс	prec	rec	F1
	(sbj+obj)	(sbj)	(sbj)	(sbj)
Baseline	75.29	00.00	00.00	00.00
Erg	86.06	50.26	50.26	50.26
TransCase(SubcatDic)	76.99	82.58	74.17	78.15
NCaseV(SubcatDic)	72.21	51.50	48.33	49.86
NCaseV(Web)	69.21	22.71	19.16	20.78
Preverbal	62.09	17.93	17.93	17.93
TransCase(Dict)	60.31	83.63	50.26	62.79
TransCase(Web)	60.10	80.94	57.47	67.21

¹We only display the features with accuracies over 60% () () () ()

Evaluation on TRAIN (crossvalidation)

Baseline and SVM system (all features line)

Feature	acc	prec(sbj)	rec(sbj)	F1(sbj)
Baseline	75.29	00.00	00.00	00.00
All features	89.62	86.34	68.89	76.63

Feature ablation results

¬SubcatDict	88.23	84.98	63.62	72.76
¬Web	88.32	83.94	65.20	73.39
¬BNC	88.23	84.49	64.14	72.93
¬Dict	87.66	86.25	59.57	70.47
¬SubcatInf	86.06	88.27	50.26	70.47
¬CaseNum	85.28	77.64	56.77	65.58
\neg NCaseV(Aux)*	87.84	83.84	62.91	71.88

メロシ メポシ メヨシ メヨシー

Evaluation on test

Evaluation against MALTParser

Results over the ambiguous relations in the test set

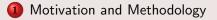
	асс	prec(sbj)	rec(sbj)	F1(sbj)
All features	89.33	82.48	71.74	76.74
MALT	86.72	76.82	65.69	70.82

Stat. significant error reduction of 19.64% (p-value <0.005).

Results over all relations in the test set

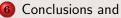
	LAS	prec(sbj)	rec(sbj)	F1(sbj)
MALT	83.17	71.57	75.01	73.24
MALT Post-processed	83.52	72.11	75.52	73.77

Stat. significant LAS improvement of 0.35 absolute points (p-value <0.00009).



- 2 Subject-object ambiguity in Basque
- 3 Features involved in the subject-object ambiguity
- 4 Experimental setup and Evaluation

6 Related work



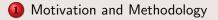
Related work

- Initial works: (Hindle & Rooth, 1993), (Ratnaparhi, 1998)
- Two main approaches to face syntactic ambiguities:
 - Enriching treebanks with additional information.
 - Parsing correction
 - Czech (Hall & NOvack, 2005)
 - German (Foth & Menzel, 2006)
 - English and Swedish (Attardi & Ciaramita, 2007)
 - Hindi (Husain et al., 2010)
 - Hindi (Husain & Agrawal, 2012)

The error reduction achieved in our work (19,64%) is considerably larger than those reported in these related works (below 10%).

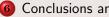
Related work

- Critic on some parse correction experiments: (Atterer and Schütze, 2007)
 - Unrealistic.
 - It relies on using the treebank as an oracle to select the ambiguous candidates.
 - Parsers do not have those gold annotations (morph and syntax) at parsing time.
- To avoid these inconveniences, when selecting candidats:
 - we used a morphological tagger
 - we used a positional heuristic for assigning dependents to verbs



- 2 Subject-object ambiguity in Basque
- B Features involved in the subject-object ambiguity
- 4 Experimental setup and Evaluation

6 Related work



6 Conclusions and future work

Conclusions and future work

- Confirmation of the relevance of complex lexical information in solving syntactic ambiguity
 - More precisely subject-object ambiguity in Basque
- All the features employed contribute positively
- The classifier obtains better results than a state-of-the art parser
- When using the output of the classifier to correct parser's output the improvement is small but statistically significant
- The most relevant features are the case and the transitivity of the verb
- Future work
 - Study the similarities and differences with typologically related languages
 - Incorporate some of the features into the treebank and statistical parser



Thank You Eskerrik asko

・ロット 「四ット 山」 シュア

33 / 33