Contribution of complex lexical information to solve syntactic ambiguity in Basque

Aitziber Atuxa    Eneko Agirre    Kepa Sarasola

Ixa Group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal pasealekua, 1 · 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.

<table>
<thead>
<tr>
<th>Ambiguity</th>
<th>English</th>
<th>Basque</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP-attachment</td>
<td>50%</td>
<td>0.01%</td>
</tr>
<tr>
<td>Subj-Obj</td>
<td>-</td>
<td>33%</td>
</tr>
</tbody>
</table>

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.*”
Methodology

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
1. Motivation and Methodology

2. Subject-object ambiguity in Basque

3. Features involved in the subject-object ambiguity

4. Experimental setup and Evaluation

5. 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:

\[
\text{absolutive} = \begin{cases} 
\text{subject of intransitive verbs} \\
\text{object of transitive verbs}
\end{cases} \\
\text{ergative} = \text{subject of transitive verbs}
\]
Subject Object ambiguity in Basque (Examples)

- Finite sentences: auxiliary marks transitivity
  1. 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

   Crisis-abs-sg finish-to hope transitive-aux-we.
   “We hope that the crisis-subj will finish. “
   ... (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."
1. Motivation and Methodology

2. Subject-object ambiguity in Basque

3. Features involved in the subject-object ambiguity

4. Experimental setup and Evaluation

5. Related work

6. Conclusions and future work
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)
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
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
Motivation and Methodology

Subject-object ambiguity in Basque

Features involved in the subject-object ambiguity

Experimental setup and Evaluation

Related work

Conclusions and future work

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

  \[
  P(\text{TransCase}) = \begin{cases} 
  \frac{\#\text{trans}}{\#\text{trans} + \#\text{intrans}} & \text{case} = \text{erg} \& P(\text{TransCase}) > 0.5 \\
  1 - P(\text{TransCase}) & \text{case} = \text{abs} \& P(\text{TransCase}) < 0.5 \\
  0 & \text{case} = \text{abs} \& P(\text{TransCase}) > 0.5 \\
  \text{none} & \text{otherwise}
  \end{cases}
  \]

- **TransCase(Web)** equivalent to **TransCase(SubcatDict)** but based on the web frequencies
Feature Space

Features related to subcategorization

- $N\text{Case}V(SubcatDict)$

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

\[
N\text{Case}V(SubcatDict) \begin{cases} 
1 & P(TransCase) > 0.5 \& P(Erg) > 0.5 \\
0 & P(TransCase) < 0.5 \& P(Erg) < 0.5 \\
\text{none} & \text{otherwise}
\end{cases}
\]

- $N\text{Case}V(Web)$ equivalent to $N\text{Case}V(SubcatDict)$ but based on the web frequencies
Feature Space

Features related to subcategorization

- \( N\text{CaseVAux}(\text{SubcatDict}) \)

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

\[
N\text{CaseVAux}(\text{SubcatDict}) \begin{cases} \frac{(n+\text{abs}+v+\text{intransAux})+(n+\text{erg}+v+\text{transAux})}{(n+\text{case}+v)} & \#(n + \text{case} + v) > 0 \\ \text{none} & \text{otherwise} \end{cases}
\]

- \( N\text{CaseVAux}(\text{Web}) \) equivalent to \( N\text{CaseVAux}(\text{SubcatDict}) \) but based on the web frequencies
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:

\[
\text{Subj(BNC)} = \begin{cases} 
1 & \text{Prob}(\text{elementTranslation} + \text{subj}) > 0 \\
0 & \text{Prob}(\text{elementTranslation} + \text{obj}) > 0 \\
\text{none} & \notin \text{BNC} \land \neg \text{translation}
\end{cases}
\]

■ **TransCase**(Dict)

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

\[
\text{TransCase(Dict)} = \begin{cases} 
1 & \text{erg} + (\text{du}|\text{du} − \text{dio}) \land \text{abs} + (\text{da}|\text{da} − \text{zaio}|\text{zaio}) \\
0 & \text{abs} + (\text{du}|\text{du} − \text{dio}) \\
\text{none} & \text{otherwise} (\text{du} − \text{da})
\end{cases}
\]
1 Motivation and Methodology

2 Subject-object ambiguity in Basque

3 Features involved in the subject-object ambiguity

4 Experimental setup and Evaluation

5 Related work

6 Conclusions and future work
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:

- 1st 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.
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 own\(^1\):

<table>
<thead>
<tr>
<th>Feature</th>
<th>acc (sj+obj)</th>
<th>prec (sj)</th>
<th>rec (sj)</th>
<th>F1 (sj)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>75.29</td>
<td>00.00</td>
<td>00.00</td>
<td>00.00</td>
</tr>
<tr>
<td>Erg</td>
<td>86.06</td>
<td>50.26</td>
<td>50.26</td>
<td>50.26</td>
</tr>
<tr>
<td>TransCase(SubcatDic)</td>
<td>76.99</td>
<td>82.58</td>
<td>74.17</td>
<td>78.15</td>
</tr>
<tr>
<td>NCaseV(SubcatDic)</td>
<td>72.21</td>
<td>51.50</td>
<td>48.33</td>
<td>49.86</td>
</tr>
<tr>
<td>NCaseV(Web)</td>
<td>69.21</td>
<td>22.71</td>
<td>19.16</td>
<td>20.78</td>
</tr>
<tr>
<td>Preverbal</td>
<td>62.09</td>
<td>17.93</td>
<td>17.93</td>
<td>17.93</td>
</tr>
<tr>
<td>TransCase(Dict)</td>
<td>60.31</td>
<td>83.63</td>
<td>50.26</td>
<td>62.79</td>
</tr>
<tr>
<td>TransCase(Web)</td>
<td>60.10</td>
<td>80.94</td>
<td>57.47</td>
<td>67.21</td>
</tr>
</tbody>
</table>

\(^1\)We only display the features with accuracies over 60%
## Evaluation on TRAIN (crossvalidation)

### Baseline and SVM system (all features line)

<table>
<thead>
<tr>
<th>Feature</th>
<th>acc</th>
<th>prec(sbj)</th>
<th>rec(sbj)</th>
<th>F1(sbj)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>75.29</td>
<td>00.00</td>
<td>00.00</td>
<td>00.00</td>
</tr>
<tr>
<td>All features</td>
<td><strong>89.62</strong></td>
<td><strong>86.34</strong></td>
<td><strong>68.89</strong></td>
<td><strong>76.63</strong></td>
</tr>
</tbody>
</table>

### Feature ablation results

<table>
<thead>
<tr>
<th>¬Feature</th>
<th>acc</th>
<th>prec(sbj)</th>
<th>rec(sbj)</th>
<th>F1(sbj)</th>
</tr>
</thead>
<tbody>
<tr>
<td>¬SubcatDict</td>
<td>88.23</td>
<td>84.98</td>
<td>63.62</td>
<td>72.76</td>
</tr>
<tr>
<td>¬Web</td>
<td>88.32</td>
<td>83.94</td>
<td>65.20</td>
<td>73.39</td>
</tr>
<tr>
<td>¬BNC</td>
<td>88.23</td>
<td>84.49</td>
<td>64.14</td>
<td>72.93</td>
</tr>
<tr>
<td>¬Dict</td>
<td>87.66</td>
<td>86.25</td>
<td>59.57</td>
<td>70.47</td>
</tr>
<tr>
<td>¬SubcatInf</td>
<td>86.06</td>
<td>88.27</td>
<td><strong>50.26</strong></td>
<td>70.47</td>
</tr>
<tr>
<td>¬CaseNum</td>
<td><strong>85.28</strong></td>
<td><strong>77.64</strong></td>
<td>56.77</td>
<td><strong>65.58</strong></td>
</tr>
<tr>
<td>¬NCaseV(Aux)*</td>
<td>87.84</td>
<td>83.84</td>
<td>62.91</td>
<td>71.88</td>
</tr>
</tbody>
</table>
Evaluation on test

Evaluation against MALTParser

- Results **over the ambiguous relations** in the test set

<table>
<thead>
<tr>
<th></th>
<th>acc</th>
<th>prec(sbj)</th>
<th>rec(sbj)</th>
<th>F1(sbj)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>89.33</td>
<td>82.48</td>
<td>71.74</td>
<td>76.74</td>
</tr>
<tr>
<td>MALT</td>
<td>86.72</td>
<td>76.82</td>
<td>65.69</td>
<td>70.82</td>
</tr>
</tbody>
</table>

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

- Results **over all relations** in the test set

<table>
<thead>
<tr>
<th></th>
<th>LAS</th>
<th>prec(sbj)</th>
<th>rec(sbj)</th>
<th>F1(sbj)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALT</td>
<td>83.17</td>
<td>71.57</td>
<td>75.01</td>
<td>73.24</td>
</tr>
<tr>
<td>MALT Post-processed</td>
<td>83.52</td>
<td>72.11</td>
<td>75.52</td>
<td>73.77</td>
</tr>
</tbody>
</table>

*Stat. significant LAS improvement of 0.35 absolute points (p-value < 0.00009).*
1. Motivation and Methodology

2. Subject-object ambiguity in Basque

3. Features involved in the subject-object ambiguity

4. Experimental setup and Evaluation

5. Related work

6. Conclusions and future 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 candidates:
  - we used a morphological tagger
  - we used a positional heuristic for assigning dependents to verbs
Motivation and Methodology

Subject-object ambiguity in Basque

Features involved in the subject-object ambiguity

Experimental setup and Evaluation

Related work

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