Evaluating Distributional Properties of Tagsets

Markus Dickinson and Charles Jochim
Department of Linguistics, Indiana University
Institute for Natural Language Processing, Universität Stuttgart

LREC
20 May 2010
Valletta, Malta
Introduction

A problem: no standard set of categories (i.e. part-of-speech tags) for evaluating category induction

- smaller mapped tagsets are often used (Goldwater and Griffiths 2007; Toutanova and Johnson 2008)
Introduction

A problem: no standard set of categories (i.e. part-of-speech tags) for evaluating category induction

- smaller mapped tagsets are often used (Goldwater and Griffiths 2007; Toutanova and Johnson 2008)

How do we evaluate tagset mappings?

*Internal quality* whether tagset can be used to tag accurately

*External quality* whether tagset captures desired linguistic phenomena

- Generally trying to capture distribution
Introduction

A problem: no standard set of categories (i.e. part-of-speech tags) for evaluating category induction

- smaller mapped tagsets are often used (Goldwater and Griffiths 2007; Toutanova and Johnson 2008)

How do we evaluate tagset mappings?

*Internal quality* whether tagset can be used to tag accurately

*External quality* whether tagset captures desired linguistic phenomena

- Generally trying to capture distribution

**Goal:** understand & evaluate the distributional properties that mappings encode
Q1: How do we measure distributional properties in tagsets?

- POS tags encapsulate some combination of morphological & syntactic (& other?) properties
Q1: How do we measure distributional properties in tagsets?

- POS tags encapsulate some combination of morphological & syntactic (& other?) properties

A1: Use tagset mappings to isolate distribution

- Can factor out morphological properties to examine only distributional
Tagsets & Tagset Mappings

**Q1:** How do we measure distributional properties in tagsets?

- POS tags encapsulate some combination of morphological & syntactic (& other?) properties

**A1:** Use tagset mappings to isolate distribution

- Can factor out morphological properties to examine only distributional

NB: Work on learner language advocates separate distributional, morphological, & lexicon tags (Díaz-Negrillo et al. 2010, to appear)
Frequent frames

Q2: What method can be used to test distribution?
Frequent frames

Q2: What method can be used to test distribution?

A2: Frequent frames distinguish distributional properties

► *Frame* = two words around a target word (Mintz 2003)

► e.g., frame *you it* generally predicts a verbal category for the target

► *Frequent* frames can be used for basic distributional grouping
Frequent frames

Q2: What method can be used to test distribution?

A2: Frequent frames distinguish distributional properties

- **Frame** = two words around a target word (Mintz 2003)
  - e.g., frame *you → it* generally predicts a verbal category for the target

- Frequent frames can be used for basic distributional grouping

Advantages of using frequent frames:

- simple to encode
- purely distributional, i.e., test nothing else
- cross linguistic, i.e., can work for different languages (Chemla et al. 2009; Xiao et al. 2006)
Lexical evaluation

Q3: What external criteria indicate the (loss in) quality of a distributional mapping?
Lexical evaluation

Q3: What external criteria indicate the (loss in) quality of a distributional mapping?

Consider:

▶ conflating base form with non-3rd present tense verb:
  ▶ prominent ambiguity for many words, e.g., accept
▶ conflating 3rd person with non-3rd person present tense verb:
  ▶ different words: accept vs. accepts
Lexical evaluation

Q3: What *external* criteria indicate the (loss in) quality of a distributional mapping?

Consider:

▶ conflating base form with non-3rd present tense verb:
  ▶ prominent ambiguity for many words, e.g., *accept*
▶ conflating 3rd person with non-3rd person present tense verb:
  ▶ different words: *accept* vs. *accepts*

A3: Measure how many word types “lose” an ambiguity in a lexicon by using a given mapping

▶ Fewer losses are desired, as this means that words are nearly as ambiguous as they were before
Initial mappings

Started with existing tagset mappings for Penn Treebank (Smith and Eisner 2005) & SUSANNE (Brants 1997)

- Used similar mappings for other tagsets
- Mappings used to evaluate category induction (e.g., Goldwater and Griffiths 2007)
Initial mappings

Started with existing tagset mappings for Penn Treebank (Smith and Eisner 2005) & SUSANNE (Brants 1997)

- Used similar mappings for other tagsets
- Mappings used to evaluate category induction (e.g., Goldwater and Griffiths 2007)

Use purity (see Manning et al. 2008) of frame to measure accuracy

- Divide most frequent category instances among all instances

Full details in the paper ...
# Initial mapping results

<table>
<thead>
<tr>
<th>Corpus mapping</th>
<th>Frames</th>
<th>Tags</th>
<th>Purity</th>
<th>Lost amb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTB</td>
<td>98</td>
<td>45</td>
<td>79.5%</td>
<td>0</td>
</tr>
<tr>
<td>PTB-17</td>
<td>98</td>
<td>17</td>
<td>89.7%</td>
<td>2038</td>
</tr>
<tr>
<td>Bro.</td>
<td>88</td>
<td>383</td>
<td>66.3%</td>
<td>0</td>
</tr>
<tr>
<td>Bro.-17</td>
<td>88</td>
<td>18</td>
<td>84.0%</td>
<td>580</td>
</tr>
<tr>
<td>SUS.</td>
<td>102</td>
<td>425</td>
<td>38.1%</td>
<td>0</td>
</tr>
<tr>
<td>SUS.-1</td>
<td>102</td>
<td>20</td>
<td>79.1%</td>
<td>652</td>
</tr>
<tr>
<td>SUS.-2</td>
<td>102</td>
<td>61</td>
<td>75.4%</td>
<td>589</td>
</tr>
<tr>
<td>TIG.</td>
<td>58</td>
<td>155</td>
<td>82.3%</td>
<td>0</td>
</tr>
<tr>
<td>TIG.-1</td>
<td>58</td>
<td>14</td>
<td>90.5%</td>
<td>2627</td>
</tr>
<tr>
<td>TUT</td>
<td>149</td>
<td>924</td>
<td>63.5%</td>
<td>0</td>
</tr>
<tr>
<td>TUT-1</td>
<td>149</td>
<td>16</td>
<td>89.6%</td>
<td>183</td>
</tr>
<tr>
<td>TUT-2</td>
<td>149</td>
<td>94</td>
<td>84.2%</td>
<td>64</td>
</tr>
</tbody>
</table>

**Table:** Original & (coarsely) mapped tag purity
Defining noun and verb mappings

- Merge nouns and verbs along two dimensions:
  - Common syntactic/semantic properties
  - Common morphological properties
Defining noun and verb mappings

- Merge nouns and verbs along two dimensions:
  - Common syntactic/semantic properties
  - Common morphological properties

Merge nouns by:

- *noun type*: pronoun (PRP), common (NN/NNS), proper (NNP/NNPS)
- *noun form*: pronoun (PRP), singular (NN/NNP), plural (NNS/NNPS)
Defining noun and verb mappings

- Merge nouns and verbs along two dimensions:
  - Common syntactic/semantic properties
  - Common morphological properties

Merge nouns by:

- **noun type**: pronoun (PRP), common (NN/NNS), proper (NNP/NNPS)
- **noun form**: pronoun (PRP), singular (NN/NNP), plural (NNS/NNPS)

Merge verbs by:

- **finiteness**: modal (MD), finite (VBP/VBZ/VBD), non-finite (VB/VBG/VBN)
- **verb form**: modal (MD), base (VB/VBP), -ed (VBD/VBN), -ing (VBG), -s (VBZ)
Penn Treebank results

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Tags</th>
<th>Purity</th>
<th>Lost amb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTB-17</td>
<td>17</td>
<td>89.7%</td>
<td>2038</td>
</tr>
<tr>
<td>N. form/V. form</td>
<td>41</td>
<td>83.2%</td>
<td>2653</td>
</tr>
<tr>
<td>N. type/V. form</td>
<td>41</td>
<td>84.3%</td>
<td>2101</td>
</tr>
<tr>
<td>N. form/Finite</td>
<td>39</td>
<td>85.1%</td>
<td>905</td>
</tr>
<tr>
<td><strong>N. type/Finite</strong></td>
<td>39</td>
<td><strong>86.3%</strong></td>
<td>352</td>
</tr>
<tr>
<td>No mappings</td>
<td>45</td>
<td>79.5%</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Results for Penn Treebank

- Noun type and verb finiteness results in high purity
Penn Treebank results

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Tags</th>
<th>Purity</th>
<th>Lost amb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTB-17</td>
<td>17</td>
<td>89.7%</td>
<td>2038</td>
</tr>
<tr>
<td>N. form/V. form</td>
<td>41</td>
<td>83.2%</td>
<td>2653</td>
</tr>
<tr>
<td>N. type/V. form</td>
<td>41</td>
<td>84.3%</td>
<td>2101</td>
</tr>
<tr>
<td>N. form/Finite</td>
<td>39</td>
<td>85.1%</td>
<td>905</td>
</tr>
<tr>
<td><strong>N. type/Finite</strong></td>
<td>39</td>
<td><strong>86.3%</strong></td>
<td>352</td>
</tr>
<tr>
<td>No mappings</td>
<td>45</td>
<td>79.5%</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table:** Results for Penn Treebank

- Noun type and verb finiteness results in high purity
  - ... while best maintaining distinctions in the lexicon
- Note that purity and lost ambiguity vary dramatically even though mapped tagsets are nearly the same size
Brown results

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Tags</th>
<th>Purity</th>
<th>Lost amb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bro.-17</td>
<td>18</td>
<td>84.0%</td>
<td>580</td>
</tr>
<tr>
<td>N. form/V. form</td>
<td>59</td>
<td>72.0%</td>
<td>1685</td>
</tr>
<tr>
<td>N. type/V. form</td>
<td>58</td>
<td>79.1%</td>
<td>1611</td>
</tr>
<tr>
<td>N. form/Finite</td>
<td>57</td>
<td>73.4%</td>
<td>188</td>
</tr>
<tr>
<td>N. type/Finite</td>
<td>56</td>
<td>80.5%</td>
<td>114</td>
</tr>
<tr>
<td>No mappings</td>
<td>383</td>
<td>66.3%</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Results for Brown

- Noun type and verb finiteness again return the highest purity and the least number of ambiguities lost.
Interlude: Issues in some tagset mappings

Some tagsets are difficult to map because they emphasize lexical properties over morphological or distributional
Interlude: Issues in some tagset mappings

Some tagsets are difficult to map because they emphasize lexical properties over morphological or distributional

- **SUSANNE**
  1. **NNc**: nouns that can be singular or plural (e.g., *sheep*)
    - Prohibits accurate mappings for singular vs. plural nouns: NNc does not properly fit into either category
  2. No distinction between base form verbs and present tense verbs (non-3rd person)
    - Prohibits accurate mapping for verb finiteness
Interlude: Issues in some tagset mappings

Some tagsets are difficult to map because they emphasize lexical properties over morphological or distributional

▶ SUSANNE

1. **NNc**: nouns that can be singular or plural (e.g., *sheep*)
   ▶ Prohibits accurate mappings for singular vs. plural nouns: NNc does not properly fit into either category

2. No distinction between base form verbs and present tense verbs (non-3rd person)
   ▶ Prohibits accurate mapping for verb finiteness

▶ Turin University Treebank (TUT)

1. Nouns that can be either singular or plural (i.e. *città*) are marked **ALLVAL** for number

2. Nouns that can be either gender (i.e. *Albanese*) are marked **ALLVAL** for gender
SUSANNE results

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Tags</th>
<th>Purity</th>
<th>Lost amb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>First letter</td>
<td>20</td>
<td>79.1%</td>
<td>652</td>
</tr>
<tr>
<td>Two letters</td>
<td>61</td>
<td>75.4%</td>
<td>589</td>
</tr>
<tr>
<td>N. form/V. form</td>
<td>279</td>
<td>67.3%</td>
<td>532</td>
</tr>
<tr>
<td>N. type/V. form</td>
<td>279</td>
<td>73.9%</td>
<td>533</td>
</tr>
<tr>
<td>N. form/Finite</td>
<td>277</td>
<td>68.4%</td>
<td>104</td>
</tr>
<tr>
<td>N. type/Finite</td>
<td>277</td>
<td>75.0%</td>
<td>105</td>
</tr>
<tr>
<td>No mappings</td>
<td>425</td>
<td>38.1%</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Results for SUSANNE

- Despite inexact mappings, results still favor noun type and verb finiteness
SUSANNE results

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Tags</th>
<th>Purity</th>
<th>Lost amb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>First letter</td>
<td>20</td>
<td>79.1%</td>
<td>652</td>
</tr>
<tr>
<td>Two letters</td>
<td>61</td>
<td>75.4%</td>
<td>589</td>
</tr>
<tr>
<td>N. form/V. form</td>
<td>279</td>
<td>67.3%</td>
<td>532</td>
</tr>
<tr>
<td>N. type/V. form</td>
<td>279</td>
<td>73.9%</td>
<td>533</td>
</tr>
<tr>
<td>N. form/Finite</td>
<td>277</td>
<td>68.4%</td>
<td>104</td>
</tr>
<tr>
<td><strong>N. type/Finite</strong></td>
<td><strong>277</strong></td>
<td><strong>75.0%</strong></td>
<td><strong>105</strong></td>
</tr>
<tr>
<td>No mappings</td>
<td>425</td>
<td>38.1%</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Results for SUSANNE

- Despite inexact mappings, results still favor noun type and verb finiteness
- Possible to have a rich tagset (e.g., 277 tags) without sacrificing accuracy
TUT results

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Tags</th>
<th>Purity</th>
<th>Lost amb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>“syntactic categories”</td>
<td>16</td>
<td>89.6%</td>
<td>183</td>
</tr>
<tr>
<td>Chanev mapping</td>
<td>94</td>
<td>84.2%</td>
<td>64</td>
</tr>
<tr>
<td>N. form/V. form</td>
<td>284</td>
<td>75.7%</td>
<td>62</td>
</tr>
<tr>
<td>N. type/V. form</td>
<td>277</td>
<td>84.5%</td>
<td>71</td>
</tr>
<tr>
<td>N. form/Finite</td>
<td>269</td>
<td>77.1%</td>
<td>63</td>
</tr>
<tr>
<td><strong>N. type/Finite</strong></td>
<td>262</td>
<td><strong>85.8%</strong></td>
<td>72</td>
</tr>
<tr>
<td>No mappings</td>
<td>924</td>
<td>63.5%</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Results for TUT

- Italian’s more complex morphology makes it difficult to use mapping by form
Automatic tag mappings

So far: we have mapped tagsets based on what we suspected were useful properties

- With large/unfamiliar tagsets, this approach can be time-consuming
- It might be helpful to have some automatic, bottom-up help in defining a mapping
Automatic tag mappings

So far: we have mapped tagsets based on what we suspected were useful properties

- With large/unfamiliar tagsets, this approach can be time-consuming
- It might be helpful to have some automatic, bottom-up help in defining a mapping

Approach:
- Use similarity measure to find & group tags that appear in the same frame contexts
  - e.g., Tags VV0t and VV0v may be mapped if they occur often as the target of the frame *he ___ to*
Using cosine similarity

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Tags</th>
<th>Purity</th>
<th>Lost amb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>First letter</td>
<td>20</td>
<td>79.1%</td>
<td>652</td>
</tr>
<tr>
<td>Two letters</td>
<td>61</td>
<td>75.4%</td>
<td>589</td>
</tr>
<tr>
<td>N. type/Finite</td>
<td>277</td>
<td>75.0%</td>
<td>105</td>
</tr>
<tr>
<td><strong>Cosine sim.</strong></td>
<td>326</td>
<td>73.3%</td>
<td>36</td>
</tr>
<tr>
<td>No mappings</td>
<td>425</td>
<td>38.1%</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Cosine similarity results for SUSANNE

Take-home points:

- Cosine similarity provides a bottom-up approach to group tags based strictly on distributional properties
- Could be a useful first step in tagset design in order to make a tagset that captures distributional properties
  - cf. also clustering methods (Miller et al. 2004)
Conclusions

1. Using frequent frames, or similar purely distributional tests, allows one to test how distributional a tagset is.

2. When evaluating POS tagging or category induction methods involving mapping to simpler tagset, one should report a measurement of external quality.
   - We propose one which records the number of ambiguities lost in the lexicon.

3. Tagset mappings can integrate both top-down linguistic knowledge and bottom-up evidence.
We wish to express our thanks to:

- The IU Computational Linguistics discussion group and the University of Stuttgart’s IfNLP Institutversammlung for helpful feedback
- Three anonymous reviewers for their useful comments and different perspectives
References


Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze (2008). Introduction to Information Retrieval. CUP.


Toutanova, Kristina and Mark Johnson (2008). A Bayesian LDA-based Model