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

Tagsets & Tagset Mappings

Q1: How do we measure distributional properties in tagsets?

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

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

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

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

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

Mapping	Tags	Purity	Lost amb.
PTB-17	17	89.7%	2038
N. form/V. form	41	83.2%	2653
N. type/V. form	41	84.3%	2101
N. form/Finite	39	85.1%	905
N. type/Finite	39	**86.3%**	**352**
No mappings	45	79.5%	0

Table: Results for Penn Treebank

- Noun type and verb finiteness results in high purity

Penn Treebank results

Mapping	Tags	Purity	Lost amb.
PTB-17	17	89.7%	2038
N. form/V. form	41	83.2%	2653
N. type/V. form	41	84.3%	2101
N. form/Finite	39	85.1%	905
N. type/Finite	39	**86.3%**	**352**
No mappings	45	79.5%	0

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

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

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

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

Table: Results for SUSANNE

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

SUSANNE results

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

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

Mapping	Tags	Purity	Lost amb.
"syntactic categories"	16	89.6%	183
Chanev mapping	94	84.2%	64
N. form/V. form	284	75.7%	62
N. type/V. form	277	84.5%	71
N. form/Finite	269	77.1%	63
N. type/Finite	**262**	**85.8%**	**72**
No mappings	924	63.5%	0

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

Mapping	Tags	Purity	Lost amb.
First letter	20	79.1%	652
Two letters	61	75.4%	589
N. type/Finite	277	75.0%	105
Cosine sim.	326	73.3%	36
No mappings	425	38.1%	0

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

Acknowledgements

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

Brants, Thorsten (1997). Internal and External Tagsets in Part-of-Speech Tagging. In *Proceedings of Eurospeech*. Rhodes, Greece.

Chemla, E., T. H. Mintz, S. Bernal and A. Christophe (2009). Categorizing words using 'Frequent Frames': What cross-linguisic analyses reveal about core principles. *Developmental Science* 12(3), 396–406.

Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera and Holger Wunsch (2010, to appear). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* .

Goldwater, Sharon and Tom Griffiths (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL-07*. Prague.

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

Miller, Scott, Jethran Guinness and Alex Zamanian (2004). Name Tagging with Word Clusters and Discriminative Training. In *Proceedings of HLT-NAACL 2004*. Boston, MA.

Mintz, Toben H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90, 91–117.

Redington, Martin, Nick Chater and Steven Finch (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science* 22(4), 425–469.

Smith, Noah A. and Jason Eisner (2005). Contrastive Estimation: Training Log-Linear Models on Unlabeled Data. In *Proceedings of ACL'05*. Ann Arbor, MI.

Toutanova, Kristina and Mark Johnson (2008). A Bayesian LDA-based Model for Semi-Supervised Part-of-speech Tagging. In *Proceedings of NIPS 2008*. Vancouver.

Xiao, L., X. Cai and T. Lee (2006). The development of the verb category and verb argument structures in Mandarin-speaking children before two years of age. In *The Seventh Tokyo Conference on Psycholinguistics*. Keio University.