Cross-lingual approach

LING575

Fei Xia

Week 9: 3/4/08

Papers

- (Hana et al., 2004) J. Hana, A. Feldman, and C. Brew. A resource-light approach to Russian morphology: Tagging Russian using Czech resources.

- (Feldman et al., 2006) A. Feldman, J. Hana, and C. Brew. Experiments in morphological annotation transfer.

(Hana et al., 2004)

Main ideas

- Use a trigram model for POS tagging
- Borrow transition model from a closely related language.
- Use a morphological analyzer to create a “lexicon”.
- Use the lexicon to create “uniform” emission probability

The Setting

Input:

- Labeled data of Czech
- Unlabeled data of Russian
- A morphological analyzer for Russian

Output:

- A Russian POS tagger

Russian and Czech

- The languages share many linguistic properties:
  - Both are Slavic languages
  - Both have extensive morphology
  - Both have free word order
  - The word order in both is very similar.

=> One can borrow transition probability from Czech and use it for Russian

The Russian morphological analyzer

- Using about 80 paradigms
- Given a word, find all possible pref + stem + suf decompositions
- Reducing ambiguity by some heuristics
  - Longest ending filtering (LEF)
  - Relations between words: e.g., talking, talked

Morphological analysis results

LEF	no	no	no	yes	yes	yes
Lexicon based on	0	100K	1M	0	100K	1M
recall	**95.4**	94	93.1	84.4	88.3	90.4
avg ambig (tag/word)	10.9	7.0	4.7	4.1	3.5	**3.1**
Tagging – accuracy	50.7	62.1	67.5	62.1	66.8	69.4

➔ The analyzer provides a (word, tag) lexicon

➔The two methods reduce the average number of tags/word from 10.9 to 3.1

Further improvement

- Transition probability: to make Czech more like Russian through preprocessing
  - Applying simple russification rules
  - Results: accuracy improves from 68% to 69.4%.

- Handling a large tagset: to train subtaggers and combine their results

The tagset for the two languages

No.	Description	Abbr.	No. of values	
			Cz	Ru
1	POS	P	12	12
2	SubPOS – detailed POS	S	75	32
3	Gender	g	11	5
4	Number	n	6	4
5	Case	c	9	8
6	Possessor's Gender	G	5	4
7	Possessor's Number	N	3	3
8	Person	p	5	5
9	Tense	t	5	5
10	Degree of comparison	d	4	4
11	Negation	a	3	3
12	Voice	v	3	3
13	Unused		1	1
14	Unused		1	1
15	Variant, Style	V	10	2

Handling a large tagset

- Build subtaggers, which are trained and tested on reduced tagsets.

- Combine the results of subtaggers by choosing the tag that agrees the most with the subtaggers.

Summary

- Model: trigram model
  - Transition probability borrowed from another language
  - Emission probability requires a lexicon, which can be created by a morphological analyzer

- Improvement:
  - Two methods to reduce (word, tag) ambiguity:
    - Baseline (random pick): 33.6%
    - HMM: 69.4%
  - Russification to improve transition probability: 69.4% to 72.6%
  - Train subtaggers and combine their results, as a way to deal with a large tagset: 72.6% to 73.5%

(Feldman et al., 2006)

Main ideas

- To test the effect of using borrowed transition probability and uniform emission probability
- To improve transition probability
- To improve the emission probability by identifying cognates.

The effect of different transition probabilities

Different ways to get transition probability (Results are from Table 2):

- From Czech labeled data: 78.6
- From Czech and Polish labeled data
  - Merging the data: 79.7
  - Interpolating the two models: 79.1
- From Russian labeled data: 81.2

The effect of different emission probabilities

Same transition probability (from Czech data), but different emission probability:

- Uniform emission probability (based on the lexicon): 78.6
- Russian emission probability (using labeled Russian data): 95.6

Using cognates

- Identify cognates in Czech and Russian
- Use the cognates to set the emission probability for Russian
- This approach provides decent improvement.