A Trivial Method for Choosing the Right Lemma

Heiki-Jaan Kaalep, Riin Kirt, Kadri Muischnek

University of Tartu

Problem area: corpus and lexicon

- What is the vocabulary of a text (or a corpus)?
- What words (=lemmas) are used?

lemma (=base form, citation form, lexicon headword):

  - sg nom for nouns, adjectives, pronouns
  - infinitive for verbs

NB! Non-changeable words pose no problem (always the same lemma)

Plan

- Lemma ambiguity in standard corpus processing workflow
- Rule-based, regular morphology of Estonian
- Text-based evidence for lemma disambiguation

Standard corpus processing workflow

1. Pre-processing (segmentation etc)

   Winds Jäägi

2. Morphological analysis (incl. guessing)

   wind + s N [wind] ‘air in motion’ jää + gi N sg nom ‘ice’
   V [wind] ‘ventilate’ jää + gi N sg gen ‘ice’
   V [wʌind] ‘twist’ jääk + i N sg gen ‘remainder’
   N [wʌind] ‘act of twisting’ jää + gi V ‘stay’
   Winds N proper Jäägi N proper

3. POS disambiguation (≈ morphological disamb.)

   wind + s N [wind] ‘air in motion’ jää + gi N sg gen ‘ice’
   N [wʌind] ‘act of twisting’ jääk + i N sg gen ‘remainder’

4. WSD (≈ lemma disambiguation)

   wind + s N [wʌind] ‘act of twisting’ jääk + i N sg gen ‘remainder’

Agglutinative, inflective language: morphological disambiguation = POS + lemma disambiguation?

wordform = lemma + morphosyntactic tags

1 / many   1 / many

kuus	kuu+s	N sg iness	‘month’	left	left	Adj
	kuus	Card sg nom	‘six’		leave	V
jää	jää	N sg nom	‘ice’	sheep	sheep	N sg
		N sg gen	‘ice’			N pl
jäägi	jää+gi	N sg nom	‘ice’	winds	wind	N [wind]
		N sg gen	‘ice’			N [wʌind]
		V	‘stay’			V [wind]
	jääk+i	N sg gen	‘remainder’			V [wʌind]

Estonian corpus tagging

1. Find all the possible morphological analyses and lemmas of all the words; guess if necessary
2. Morphological disambiguation. Choose the most likely analyses, based on the sentential context (grammatical tag sequence)

⇒ 1.5% tokens with unique tags, multiple lemmas

⇒ Liisiga

Liis + ga	N prop sg komitative
Liisi + ga	N prop sg komitative

jäägi

jää + gi	N sg gen
jääk + i	N sg gen

Sources of lexical ambiguity

1. Homonymous case forms in the dictionary

   jäägi => jää+gi / jääk+i   sg gen   ‘ice / remainder’

   teod => tegu+d / tigu+d pl nom ‘deeds / snails’

2. Guessed lemmas from word forms

   ...erlaadid => ...erlaat+d / ...erlaad+d pl nom

3. Guessed proper noun lemmas

   Liisi => Liis+i / Liisi

   Liisiga => Liis+ga / Liisi+ga

4. Parallel forms in singular nominative in the dictionary

   päikene = päike ‘sun’, manner = mander ‘continent’

Rule-based guessing is based on regular inflection (simplex words)

CVCCV ratsu, CVVCV Liisi

sg nom	sg gen	sg part	sg illative	pl nom	pl part
ratsu	ratsu	ratsut	ratsusse	ratsude	ratsusid
X	X	X+t	X+sse	X+de	X+sid

Liisi, Liisi, Liisit, Liisisse, Liiside, Liisisid

VVC siid, Liis, VCC[C] link

link	lingi	linki	linki	linkide	linke
X_S	X_W+ i	X_S+ i	X_S+ i	X_S+ ide	X_S+ e

Liis, Liisi, Liisi, Liisi, Liiside, Liise

How to choose the right lemma?

First idea: probability

... Where would you get the prob estimation from?

Corpus similarity problem (ice /remainder)

You haven' t seen OOV (incl. names) before...

How to choose the right lemma?

Idea: look at a wider context, perhaps there is some evidence?

Algorithm

1. Make a frequency list LL of all the lemmas (in this text); if a word has multiple lemmas, include them all (both Liis and Liisi)
2. For every token with multiple lemmas, keep only the most frequent one from LL

(Note this is quite opportunistic)

Example

1. Text:

   ... ... Liisiga ... ... Liisit ... ...

2. Analyzed and disambiguated:

Liisiga	=> Liis + ga / Liisi + ga
Liisit	=> Liisi + t

3. Frequency list LL:

   Liis 1, Liisi 2

4. Lemma chosen:

   Liisiga => Liisi + ga

Evaluation

110,000 tokens (fiction, newspaper, science)

	Total	Ambiguous from the dictionary	Ambiguous from the guesser	Proper names from the guesser	Parallel forms in the nominative
Initial lemma ambiguity	1670	620	280	640	130
Disamb-ed	1190	530	220	340	100
Correct	1120	500	190	330	100
Erroneous	80	30	40	10	0
Unchanged	480	90	60	300	30
Precision	0.94	0.94	0.86	0.97	1.0
Recall	0.67	0.81	0.68	0.52	0.77

Why does it work?

- Thanks to data sparcity.
- One sense per discourse!
  - But how do you define discourse? What is the right text span?
  - Go from smaller texts to larger; several iterations
- Rule-based, consistent guesser
  - But if the sentential disambiguator makes an error, it may ruin the frequency statistics
- Worry about evidence, not probability

Final thoughts

- Real texts require tools that do not rely on lexicons only
- Looking beyond sentence is easy and useful
  - Used the method for tagging 200 M corpus.
- Use the evidence, don' t guess!
- How language-specific is this approach? Is the lemma ambiguity a complete non-problem for, say, Latvian?