GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Sentence selection

in the context of

Swedish as a second language:

potential for GDEX

EUROPEAN NETWORK OF e-LEXICOGRAPHY

Elena Volodina & Ildikó Pilán

COST ENeL WG workshop

Vienna, 12 February 2015

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Overview

- Introduce the initial article
- Pass on to the tests combining machine learning and heuristics
- Present the present-day state in research for Swedish

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Semi-automatic selection of best corpus examples for Swedish

- by Elena Volodina, Richard Johansson, Sofie Johansson Kokkinakis
- published at a workshop NLP4CALL, in 2012

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Background

- Selection of examples for L2 training and Lexicography:
  - Invent – subjective and time-consuming
  - Select manually – hundreds of corpus hits, time-consuming
  - (Semi-)automatic pre-selection – a possible alternative

- Principle: rank examples according to their appropriateness or “goodness”; the best ones come to the top

- Definition of “goodness” in linguistic parameters:
  - Optimal sentence length
  - Optimal word length
  - Presence of subject and finite verb
  - etc.
- Previous tests with automatic ranking: for English (Kilgariff et.al. 2008), for Slovene (Kosem et.al. 2011), for German (Segler 2007, Didakowski et.al. 2012)

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Ranking algorithms for Swedish

- **Algorithm #1** (manually defined rules)
  - Each example scored independently using set of heuristic rules with associated weights
  - Sentence length, word frequency, keyword position, presence of a finite verb
  - Only “soft” parameters, i.e. points withdrawn, examples are considered anyway through their ranking placement

- **Algorithm #2** (computationally calculated)
  - Principle: examples should be both typical and different (collocationally, distributionally)
  - Difference is formalized as a similarity metric based on Euclidean distance between feature vectors (words and syntactic relations)

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Evaluation set-up 1

- **Critical questions**:
  - Can the two algorithms satisfactorily rank corpus examples?
  - Which of the two performs better?
  - What parameters/predictors to consider in future development?

- **Evaluators' background**:
  - L2 teachers/computational linguists
  - Lexicographers/computational linguists
  - Lexicographer
  - All have doctoral degrees
  - 50-50 native versus non-native speakers
  - 50-50 men versus women

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Evaluation set-up 2

- **Test items:**
  - 50 test items from a graded resource (Kelly list); 10 items per proficiency level
  - Only lexical items: nouns, verbs, adjectives, adverbs
  - Nr of items per word class reflects word class distribution per proficiency level

- **Database**:
  - Three top hits per algorithm stored in a database (i.e. 6 per test item)
  - Examples selected from a combination of corpora (44,3 mln. tokens)
  - Same examples for each evaluator
  - Information about algorithm not revealed to evaluators to avoid bias

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

User interface

Select your professional group and enter a user name | Evaluation tips

Any other non-Swedish speaker | User name | user2 | Start | Resume

Your submission status

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

Evaluate corpus hits

Mark examples with symbols *acceptable*, *unacceptable* or *doubtful*. Write your comment in the text field (not obligatory). Click on the *Submit* button to save your result and get a new item.

7. resa, substantiv; cefr=A1

Nr	Corpus hits	Your rating	Your comment if any
1	Flera **resor** har stoppats av lagexperter .		
2	Under **resans** gång har åtalet justerats .		
3	**Resorna** har riksdagen betalt .		word order is not optimal for L2 learners

Acceptable

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Evaluation results. Quantitative data

	acc	*unacc*	*doubtful*	*total*
alg# 1	56,6%	19,7%	23,7%	100%
alg #2	50,3%	27%	22,7%	100%
Total (#1+#2)	53,5%	23,3%	23,1%	100%

- Alg#1 “won” by 6,3% over #2, generally
  - “well-formedness” (#1) dominates when examples are not presented as a group to demonstrate dispersion (#2)

Evaluation results. Quantitative data

user groups		*acc*	*unacc*	*doubtful*	*total*
Lexicographers, total		63,6%	20%	16,4%	100%
	alg #1	66,1%	18,6%	15,3%	100%
	alg #2	61,1%	21,4%	17,5%	100%
L2 teachers, total		46,7%	25,5%	27,7%	100%
	alg #1	50,2%	20,4%	29,3%	100%
	alg #2	43,2%	30,6%	26,1%	100%

#1 won by 5%

#1 won by 7%

- Lexicographers more positive than L2 teachers: 63,6% vs 46,7%

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Evaluation. Qualitative data.

- **Structural features to avoid:** ellipsis, passive, anaphora, pronouns, long (deep) phrase structure, non-context free sentences, unusual word order, a-typical word class patterns
- **Lexical features to avoid**: non-frequent vocabulary, proper names, acronyms, abbreviations, compounds, keyword repetition
- Criticism against **annotation errors**
- **Heterogeneous**: metaphoric use, abstract use, strange examples, etc.

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Conclusions

- **Add parameters (features)** (for rule-based heuristics)
- **Add word sense discrimination**
- Set-up a **customizable user interface** & allow users to assign weights to features for experiments
- Generate **larger output sets** (not three top examples)
- Zoom into **user group needs**, and add machine learning
- Suggest **best parameter configuration** per user group

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Conclusions

- **Sentence readability** needs to be studied
- **Need for a collection of good examples** for examination in contrast with “not-so-good” ones

Actions

- **Sentence readability** does not exist as a field – let's start it!
- A **corpus of coursebook texts** labeled by proficiency level collected (COCTAILL)
- **Lärka module** for experiments with weights added --- and more!

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Thank you!