Ranking corpus examples from "best" down

Initial user evaluation of hit-ex algorithms for Swedish

CLT

Elena Volodina, Richard Johansson, Sofie Johansson Kokkinakis
Centre for Language Technology, Språkbanken, University of Gothenburg, Sweden

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Background

- Selection of examples for L2 training and Lexicography:
  - Invent – subjective and time-consuming
  - Select manually – hundreds of corpus hits, selection becomes 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)

Ranking algorithms for Swedish (hit-ex)

- **Algorithm #1** at the moment:
  - each example is scored independently of all other examples using a manually defined set of heuristic rules, each of which has an associated weight
  - parameters under consideration: sentence length; word frequency; keyword position in the sentence; presence of finite verb;
  - only "soft" parameters, i.e. if they are not met, examples are considered anyway, through punished by withdrawing points;
  - equal "punishment" for each parameter;

- **Algorithm #2** (Borin et.al. 2012):
  - Principle: examples should not only be typical but also different
  - Difference is formalized as a similarity metric, based on the Euclidean distance between feature vectors (words and syntactic relations);
  - Vectors represent words in the context of the search terms, as well as a number of syntactic features derived from dependency trees.

Evaluation set-up 1

- Critical questions:
  - Can the two algorithms satisfactorily rank examples?
  - Which of the two performs better?
  - What other parameters might be necessary to include to improve their performance as predictors of good examples.

- Evaluators background:
  - L2 teachers/computational linguists (2)
  - lexicographers/computational linguists (2)
  - lexicographer (1)
  - all 5 have doctoral degrees

  Evaluators' mother tongues:
  - 3 native and 2 non-native Swedish speakers

  Evaluators' gender:
  - 2 men, 3 women

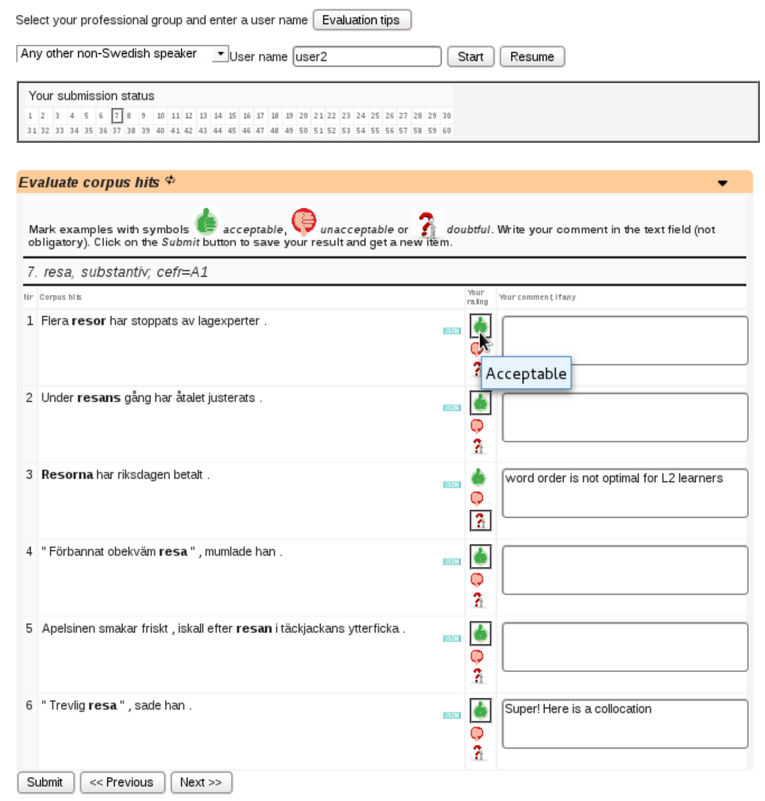
- Test items (keywords):
  - 60 test items from Kelly list, 10 per proficiency level defined in CEFR terms (Council of Europe, 2001)
  - Only lexical word classes: nouns, verbs, adjectives, adverbs
  - Number of items from each word class reflects word class distribution per proficiency level

Evaluation set-up 2

- Examples (stored in a database):
  - Three top examples per algorithm (i.e. 6 sentences per Kelly item)
  - Information about algorithm not revealed to avoid bias
  - Examples selected from a combination of corpora: SUC, Talbanken, LäsBarT, Parole, Fiction Prose, totaling at 44,3 mln. tokens)
  - Same examples for each evaluator

- Evaluators' input (stored in a database):
  - In terms of "acceptable", "unacceptable", "doubtful" per each example;
  - Plus non-obligatory comments

Evaluation. User interface



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" of examples in isolation (#1) versus dispersion better perceived in a group of examples (#2)

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

alg#1 "won" by 5% for lexicographers and by 7% for L2 teachers

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%

Totally: 54% of examples approved; 57% for alg#1 and 50% for alg#2

Evaluation. Quantitative data 2.

CEFR levels	*acc*	*unacc*	*doubtful*	*total*
A1	51,3%	27,2%	21,5%	100%
alg #1	49%	27,5%	23,5%	100%
alg #2	53,7%	26,8%	19,5%	100%
A2	48,7%	20,7%	30,7%	100%
alg #1	57,3%	12%	30,7%	100%
alg #2	40%	29,3%	30,7%	100%
B1	47,7%	31,3%	21%	100%
alg #1	56%	22,7%	21,3%	100%
alg #2	39,3%	40%	20,7%	100%

alg#1 outperforms alg #2 for levels A2, B1, B2

Possible reasons:
- individual well-formedness is important for beginner (A2) and intermediate (B1, B2) levels

alg#1 and #2 perform almost equally for levels A1, C1, C2

CEFR levels	*acc*	*unacc*	*doubtful*	*total*
B2	58,3%	18,7%	23%	100%
alg #1	60,7%	16,7%	22,7%	100%
alg #2	56%	20,7%	23,3%	100%
C1	53,7%	21%	25,3%	100%
alg #1	55,3%	19,3%	25,3%	100%
alg #2	52%	22,7%	25,3%	100%
C2	61,3%	21%	17,7%	100%
alg #1	61,3%	20%	18,7%	100%
alg #2	61,3%	22%	16,7%	100%

Possible reasons:
- for advanced (C1) and proficient (C2) levels individual well-formedness is less important

- at absolute beginner level (A1) vocabulary is easy and frequent – hence good choice of easy well-formed examples

Evaluation. Qualitative data

User comments fall into 4 categories:

1. Comments/criticism of structural features, e.g. ellipsis, passive, limited context, word order, anaphora, pronouns, long phrase structure; lack of word class specific patterns

2. Lexical features: stricter word frequency filtering, proper names, acronyms and abbreviations, compounds, keyword repetition

3. Annotation: part-of-speech specific searches, e.g. exclude proper names when searching for nouns; some annotation errors

4. Heterogeneous: metaphoric use, abstract use, strange examples, etc.

Future

1. Algorithms improvement (based on results of the evaluation):
   - (for alg #1 and #2) additional parameters, e.g. voice, word order, proper names, pronouns, strength of collocation with contextual words, valency for verbs, word class specific approaches, vocabulary frequency;
   - (for alg #2) additional techniques: word sense discrimination (Purandare and Pedersen 2004)

2. Second evaluation set-up:
   - Parameter configuration customizable in terms of strength of "punishment" per parameter
   - Larger output set for better overview (esp. for alg.#2)
   - Based on polysemous words (esp. for alg.#2)
   - More specific study over different user group needs (L2 teachers, lexicographers, linguists)

3. Potential results:
   - Suggest best parameter configuration per user group
   - Set up web service for example rating
   - Include web service into the present applications, e.g. Lärka, Korp, editing tool for Swedish FrameNet