ANNLOR: A Naïve Notation-system for Lexical Outputs Ranking

Anne-Laure Ligozat, Cyril Grouin, Anne García-Fernandez, Delphine Bernhard
{annlor,cyril.grouin}@limsi.fr, anne.garcia-fernandez@cea.fr, dbernhard@unistra.fr

Description and preparation

English Lexical Simplification task

- **task objective:** determining the degree of simplicity of words;
- **inputs:**
  - a short text in which a target word was chosen:
    <context>During the siege , George Robertson had appointed Shuja-ul-Mulk , who was a <head>bright</head> boy only 12 years old and the youngest surviving son of Aman-ul-Mulk , as the ruler of Chitral .</context>
  - several substitutes for the target word that fit the context: *intelligent;bright;clever;smart*

Corpus preprocessing

- **corpus division:** division of the trial corpus into training (66%) and evaluation (33%) sub-corpora, so as to use machine-learning based approaches;
- **corpus cleaning:** HTML entities have been replaced by their refering symbols;
- **inflection:** target words are *inflected* in the sentence while substitutes are *lemmatized*: use of Perl modules and DELA to obtain inflected forms.

Methods

Simple English Wikipedia-based system: "ANNLOR-simple"

- **hypothesis:** training a system on documents written by/for non-native English speakers would be useful: the Simple English Wikipedia targets people who do not have English as their mother tongue;
- **preparation:**
  - extraction of the textual content from the Simple English Wikipedia dump: 10 million words;
  - extraction of word n-grams with their frequencies. Number of distinct extracted n-grams:

1-gram	2-grams	3-grams	1 to 3-grams
301,718	2,517,394	6,680,906	9,500,018

- **process:** the possible substitutes of a lexical item are ranked according to the computed frequencies, in descending order:

substitutes	*intelligent*	*bright*	*clever*	*smart*
frequencies	206	475	141	201
final ranking	2	1	4	3

- **experiment:** since substitutes are lemmatized, we conducted an experiment where we lemmatized the whole corpus before counting n-grams.
- **evaluation:** the official ANNLOR-simple system used in the challenge is "*1 to 3-grams*":

reference n-grams	trial corpus	test corpus
only 1-grams	0.333	—
1 and 2-grams	0.371	—
1 to 3-grams (ANNLOR-simple)	**0.381**	**0.465**
lemmatized 1 to 3-grams	0.380	0.462

Other frequency-based methods

- **main idea:** *the more frequent a word is, the simpler it is*; new experiments on other reference corpora;
- **evaluation:** the results obtained on the trial corpus being very close from the ANNLOR-simple system, we did not use a system from these experiments in the challenge;
  - BNC corpus: score = 0.347;
  - Google Books NGrams: score = 0.367 (we only kept alphabetical 1 to 4-grams: 477,543,736 n-grams);
  - Microsoft Web N-gram Service: score = 0.383.

Contextual methods: "ANNLOR-lmbing"

- **main idea:** according to the contexts, different substitutes can be used or ranked differently;
- **process:** obtention of joint probabilities for text units from the Microsoft Web N-gram Service;
- **evaluation:** the official ANNLOR-lmbing system used in the challenge is the "*4/4 contextual window*":

left/right context size	0/3	3/0	2/2	3/3	**4/4**
score on the trial corpus	0.362	0.358	0.365	0.358	**0.370**

Combination of methods

- **main idea:** combination of each method using SVMRank (default parameters);
- **process:**
  - conversion of each output system into a feature file *(instance id, substitute, frequency, rank)*;
  - combination of all feature files after basic query-wise feature scaling.
- **evaluation:** on the evaluation sub-corpus from the trial corpus:

Simple English Wikipedia alone	Microsoft Web NGrams alone	SVMRank
0.352	0.352	0.354

Discussion and conclusion

Discussion

- on the Simple English Wikipedia, some n-grams are missing;
- quite small improvement obtained when combining Simple English Wikipedia and Microsoft NGrams.

Conclusion

- best results obtained using frequencies from the Simple English Wikipedia (ANNLOR-simple system: score = 0.381 on the trial corpus, 0.465 on the test corpus);
- we found this task hard to solve since none of our experiments outperforms the Simple Frequency baseline (score = 0.399 on the trial corpus, 0.471 on the test corpus);
- all our systems using contextual information did not achieve high scores.