ANNLOR: A Naïve Notation-system for Lexical Outputs Ranking

Anne-Laure Ligozat
LIMSI-CNRS/ENSIIE
rue John von Neumann
91400 Orsay, France
annlor@limsi.fr

Cyril Grouin
LIMSI-CNRS
rue John von Neumann
91400 Orsay, France
cyril.grouin@limsi.fr

Anne Garcia-Fernandez
CEA-LIST
NANO INNOV, Bt. 861
91191 Gif-sur-Yvette cedex, France
anne.garcia-fernandez@cea.fr

Delphine Bernhard
LiLPa, Université de Strasbourg
22 rue René Descartes, BP 80010
67084 Strasbourg cedex, France
dbernhard@unistra.fr

Abstract

This paper presents the systems we developed while participating in the first task (English Lexical Simplification) of SemEval 2012. Our first system relies on n-grams frequencies computed from the Simple English Wikipedia version, ranking each substitution term by decreasing frequency of use. We experimented with several other systems, based on term frequencies, or taking into account the context in which each substitution term occurs. On the evaluation corpus, we achieved a 0.465 score with the first system.

1 Introduction

In this paper, we present the methods we used while participating to the Lexical Simplification task at SemEval 2012 (Specia et al., 2012). We experimented with several methods:

- using word frequencies or other statistical figures from the BNC corpus, Google Books NGrams, the Simple English Wikipedia, and results from the Bing search engine (with/without lemmatization);
- using association measures for a word and its context based on language models (with/without inflection);
- making a combination of previous methods with SVMRank.

Depending on the results obtained on the training corpus, we chose the methods that seemed to best fit the data.

2 Task description

2.1 Presentation

The Lexical Simplification task aimed at determining the degree of simplicity of words. The inputs given were a short text, in which a target word was chosen, and several substitutes for the target word that fit the context.

An example of a short text follows; the target word is "outdoor", and other words of this text will be considered as the *context* of this target word.

```
<instance id="270">
  <context>With the growing demand for
      these fine garden furnishings ,
      they found it necessary to dedicate
      a portion of their business to
      <head>outdoor</head> living and
      patio furnishings .</context>
</instance>
```

The substitutes given for this target word were the following: "alfresco;outside;open-air;outdoor;". The objective was to order these words by descending simplicity.

2.2 Corpora

Two corpora were provided: the *trial* corpus with development examples, and the *test* corpus for evaluation.

In the *trial* corpus, a gold standard was also given. For the previous example, it stated that the substitutes had to be in the following order: "outdoor open-air outside, alfresco", "outdoor" being considered as the simplest substitute, and "outside" and "alfresco" being considered as the less simple ones.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 487–492,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

Three baselines have been given by the organizers: the first one is a simple randomization of the substitute list, the second one keeps the substitute list as it is, and the third one (called "simple frequency") relies on the use of the Google Web 1T corpus.

3 Preprocessing

3.1 Corpus constitution

In order to use machine-learning based approaches, we produced two sub-corpora respectively for the training and evaluation stages from the *trial* corpus. The training sub-corpus is used to develop and tune the systems we produced while the evaluation sub-corpus is used to evaluate the results of these systems.

For each set from the SemEval trial corpus, if the set is composed of at least eight lexical elements belonging to the same morpho-syntactic category (e.g., a set with at least eight instances of "bright" as an adjective), we extracted three instances from this set for the evaluation sub-corpus, the remaining instances being part of the training sub-corpus. If the set is composed of less than eight instances, all instances are used in the training sub-corpus. We also kept two complete sets of lexical elements for the evaluation sub-corpus in order to test the robustness of our methods on new lexical elements that have not been studied yet. This distribution allows us to benefit from a repartition between training and evaluation sub-corpora where the instances ratio is of 66/33%.

3.2 Corpus cleaning

While studying the trial corpus, we noticed that the texts were not always in plain text, and in particular contained HTML entities. As some of our methods used the context of target words, we decided to create a cleaner version of the corpora. For the dash and quote HTML entities (– “ etc.), we replaced each entity by its refering symbol. When replacing the apostrophe HTML entity ('), we decided to link the abbreviated token with the previous one because all n-grams methods worked better with abbreviated terms of one token-length *(don't)* than two token-length *(do n't)* (see section 5).

3.3 Inflection

In some sentences, the target words are inflected, but the substitutes are given in their lemmatized forms. For example, one of the texts was the following :

```
<context>In fact , during at least six
    distinct periods in Army history
    since World War I , lack of trust and
    confidence in senior leaders caused
    the so-called best and
    <head>brightest</head> to leave the
    Army in droves .</context>
```

For this text and target word, the proposed substitutes were "capable; most able; motivated; intelligent; bright; clever; sharp; promising", and if we want to test the simplicity of the words in context, for example with a 2-words left context, we will obtain unlikely phrases such as "best and capable" (which should be "best and most capable"). We thus used several resources to get inflected forms of words: we used the Lingua::EN::Conjugate and Lingua::EN::Inflect Perl modules, which give inflected forms of verbs and plural forms of nouns, as well as the English dictionary of inflected forms DELA,[1] to validate the Perl modules outputs if necessary, and get comparatives and superlatives of adjectives, and a list of irregular English verbs, also to validate the Perl modules outputs.

4 Simple English Wikipedia based system

Our best system, called ANNLOR-simple, is based on Simple English Wikipedia frequencies. As the challenge focused on substitutions performed by non-native English speakers, we tried to use linguistic resources that best fit this kind of data. In this way, we made the hypothesis that training our system on documents written by or written for non-native English speakers would be useful.

The use of the Simple English version from Wikipedia seems to be a good solution as it is targeted at people who do not have English as their mother tongue. Our hypothesis seems to be correct due to the results we obtained. Morevover, the Simple English Wikipedia has been used previously in work on automatic text simplification, e.g. (Zhu et al., 2010).

[1] http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html

First, we produced a plain text version of the Simple English Wikipedia. We downloaded the dump dated February 27, 2012 and extracted the textual contents using the `wikipedia2text` tool.[2] The final plaintext file contains approximately 10 million words.

We extracted word n-grams (n ranging from 1 to 3) and their frequencies from this corpus thanks to the Text-NSP Perl module [3] and its *count.pl* program, which produces the list of n-grams of a document, with their frequencies. Table 1 gives the number of n-grams produced.

Table 1: Number of distinct n-grams extracted from the Simple English Wikipedia

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

Some of these n-grams are invalid, and result from problems when extracting plain text from Wikipedia, such as "27|ufc 1", which corresponds to wiki syntax. As we would not find these n-grams in our substitution lists, we did not try to clean the n-gram data.

Then, we ranked the possible substitutes of a lexical item according to these frequencies, in descending order. For example, for the substitution list (intelligent, bright, clever, smart), the respective frequencies in the Simple English Wikipedia are (206, 475, 141, 201), and the substitutes will be ranked in descending frequencies: (bright, intelligent, smart, clever).

Several tests were conducted, with varying parameters. We used the plain text version of the Simple English Wikipedia, but also tried to lemmatize it, since substitutes are lemmatized. We used the TreeTagger [4] (Schmid, 1994) and applied it on the whole corpus, before counting n-grams. Moreover, since bigrams and trigrams increase a lot, the size of n-gram data, we evaluated their influence on results. These tests are summed up in table 2.

Table 2: Results obtained with the Simple English Wikipedia based system, on the trial and test corpora

reference n-grams	lemmas	score on trial corpus	score on test corpus
1-grams only	no	0.333	–
1 and 2-grams	no	0.371	–
1 to 3-grams	no	0.381	0.465
1 to 3-grams	yes	0.380	0.462
Simple Frequency baseline		0.398	0.471
WLV-SHEF-SimpLex (best system @SemEval2012)		–	0.496

With unigrams only, 158 substitutes of the trial corpus are absent of the reference dataset, 105 when adding bigrams, and 91 when adding trigrams. Most of the missing n-grams (when using 1 to 3-grams) indeed seem to be very uncommon, such as "undomesticated" or "telling untruths".

The small difference between the lemmatized and inflected versions of Wikipedia is due to two reasons: some substitutes are found in the lemmatized version because substitutes are given in the lemmatized form (for example "abnormal growth" is only present in its plural form "abnormal growths" in the inflected Wikipedia); and some other substitutes are missing in the lemmatized version, mostly because of errors from the TreeTagger (for example "be scared of" becomes "be scare of").

We kept the system that obtained the best scores on the trial corpus, that is with 1 to 3-grams and non-lemmatized n-grams, with a score of 0.381. This system obtained a score of 0.465 on the evaluation corpus, thus ranking second ex-aequo at the SemEval evaluation.

[2] See `http://www.polishmywriting.com/download/wikipedia2text\_rsm\_mods.tgz` and `http://blog.afterthedeadline.com/2009/12/04/generating-a-plain-text-corpus-from-wikipedia`

[3] `http://search.cpan.org/~tpederse/Text-NSP-1.25/lib/Text/NSP.pm`

[4] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

5 Other frequency-based methods

We tried several other reference corpora, always with the idea that the more frequent a word is, the simpler it is. We used the BNC corpus,[5] as well as the Google Books NGrams.[6] These NGrams were calculated on the books digitized by Google, and contain for each encountered n-gram, its number of occurrences for a given year. As the Google Books NGrams are quite voluminous, we selected a random year (2008), and kept only alphabetical n-grams with potential hyphens, and used n-grams for n ranging from 1 to 4. The dataset used contains 477,543,736 n-grams.

We also used the Microsoft Web N-gram Service (more details on this service are given in the following section) to rank substitutes in descending order. The results of these methods on the trial corpus are given in table 3. The result of the simple frequency baseline is also given: this baseline is also frequency-based, but words are ranked according to the number of hits found when querying the Google Web 1T corpus with each substitute.

Table 3: Results obtained with frequency-based methods, on the trial corpus

reference corpus	score
BNC	0.347
Google Books NGrams	0.367
Microsoft NGrams	0.383
Simple Frequency baseline	0.398

This table shows that all frequency-based methods have lower scores than the Simple Frequency baseline, although the score obtained with the Microsoft NGrams is quite close to the baseline. The results from Microsoft Ngrams and the Simple English are very close. We decided to submit the Simple English Wikipedia-based system because it was more different from the simple frequency baseline.

6 Contextual methods

We also wanted to use contextual information, since, according to the contexts of the target word, different substitutes can be used, or ranked differently. In the following two examples, the same word "film" is targetted, and the same substitutes are proposed "film;picture;movie;"; yet, in the gold standard, "film" is placed before "movie" in instance 19, and after it in instance 15.

```
<instance id="15">
  <context>Film Music Literature
      Cyberplace – Includes
      <head>film</head> reviews , message
      boards , chat room , and images
      from various films.</context>
</instance>
(...)
<instance id="19">
  <context>A fine score by George Fenton
      ( THE CRUCIBLE ) and beautiful
      photograhy by Roger Pratt add
      greatly to the effectiveness of the
      <head>film</head> .</context>
</instance>
```

Ranking substitutes thus depends on the context of the target word. We implemented two systems taking the context of target words into account.

6.1 Language model probabilities

The other system submitted (called ANNLOR-lmbing) relies on language models, which was the method used by the organizers in their Simple Frequency baseline. While the organizers used Google n-grams to rank terms to be substituted by decreasing frequency of use, we used Microsoft Web n-grams in the same way. Nevertheless, we also added the contexts of each term to be substituted.

We used the Microsoft Web N-gram Service[7] to obtain joint probability for text units, and more precisely its Python library.[8] We used the *bing-body/apr10/*) N-Gram model.

We considered a text unit composed of the lexical item and a contextual window of 4 words to the left and 4 words to the right (words being separated by spaces). For example, in the following sentence, we tested "He brings an incredibly rich and diverse background that", and the same unit with the target word replaced by substitutes, for example "He brings an incredibly lush and diverse background that".

[5] http://www.natcorp.ox.ac.uk/

[6] http://books.google.com/ngrams/datasets

[7] http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx

[8] http://web-ngram.research.microsoft.com/info/MicrosoftNgram-1.02.zip

```
<instance id="118">
  <context>He brings an incredibly
      <head>rich</head> and diverse
      background that includes everything
      from executive coaching , learning
      &amp;amp; development and management
      consulting , to senior operations
      roles , mixed with a masters in
      organizational
      development .</context>
</instance>
```

We performed several tests, with different N-Gram models, and different context sizes. Some of these results for the trial corpus are given in table 4.

Table 4: Results obtained with Microsoft Web N-gram Service, on the trial corpus

Size of left context	Size of right context	Score
0	3	0.362
3	0	0.358
2	2	0.365
3	3	0.358
4	4	0.370

For the evaluation, this system was our second run, with the parameters that obtained the best scores on the training corpus (contexts of 4 words to the left and to the right). This method obtained a 0.370 score on the trial corpus and a 0.396 score on the test corpus.[9]

7 Combination of methods

As each method seemed to have its own benefits, we tried to combine them using SVMRank [10](Joachims, 2006). The output of each system is converted into a feature file. For example, the output of the Simple English Wikipedia based system begins with:

```
1    bright    475    1
1    intelligent    206    2
1    smart    201    3
1    clever    141    4
2    light    3241    1
2    clear    707    2
2    bright    475    3
2    luminous    14    4
2    well-lit    0    5
```

The first column represent the instance id, the second one the considered substitute, the third one the feature (in this case, the frequency of the substitute in the Simple English Wikipedia), and the last one, the substitute rank according to this method. Then, we combined these files to include all features (after basic query-wise feature scaling). For example, the training file begins with:

```
1 qid:1 2:-0.00461061395325929
    3:0.0345010535723618
    #intelligent
2 qid:1 2:-0.00485010755325339
    3:-0.0213467053270483 #clever
3 qid:1 2:-0.00462903653787422
    3:0.092640777900771 #smart
4 qid:1 2:-0.00361947890097599
    3:0.0489145618699556 #bright
1 qid:4 2:-0.00461061395325929
    3:0.0345010535723618
    #intelligent
```

The first column gives the gold standard rank for the substitute (in training phase), the second one the instance id, and then feature ids and values for each substitute. Default parameters were used.

We used the division of the *trial* corpus into a training corpus and a development corpus. Table 5 gives some examples of scores obtained by combining two methods. The scores are not exactly those presented earlier, since they correspond to a part of the *trial* corpus only. Even though some improvement can be obtained by this combination, it was quite small, and so we did not use it for the evaluation.

Table 5: Results obtained with combination of methods with SVMRank, on the trial corpus

Simple English Wikipedia	Microsoft NGrams	SVM
0.352	0.352	0.354

[9]This result is different from the official one, because an incorrect file was submitted at the time.

[10]http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

8 Conclusion

In this paper, we present several systems developed for the English Lexical Simplification task of SemEval 2012. The best results are obtained using frequencies from the Simple English Wikipedia. We found the task quite hard to solve, since none of our experiments significantly outperforms the Simple Frequency baseline. On the trial corpus, our system based upon the Simple English Wikipedia achieved a score of 0.381 (below the 0.399 baseline score); on the test corpus, we achieved a score of 0.465 with the Simple English Wikipedia system while the baseline achieved a score of 0.471 score. All our systems using contextual information did not achieve high scores.

References

Thorsten Joachims. 2006. Training Linear SVMs in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. of the International Conference on New Methods in Language Processing*, Manchester, UK.

Lucia Specia, Sujay K. Jauhar, and Rada Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In *Proc. of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montréal, Canada.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1353–1361.