The LEXenstein Manual

Gustavo Henrique Paetzold

April 2016

Table of contents

Chapter 1

Introduction

LEXenstein is a framework for Lexical Simplification. It contains several methods, tools and resources for one to easily create Lexical Simplification systems and test them in various distinct ways. It takes Lexical Simplification as a process that can be modeled as a pipeline of sub-processes, as illustrated in Figure 1.1.

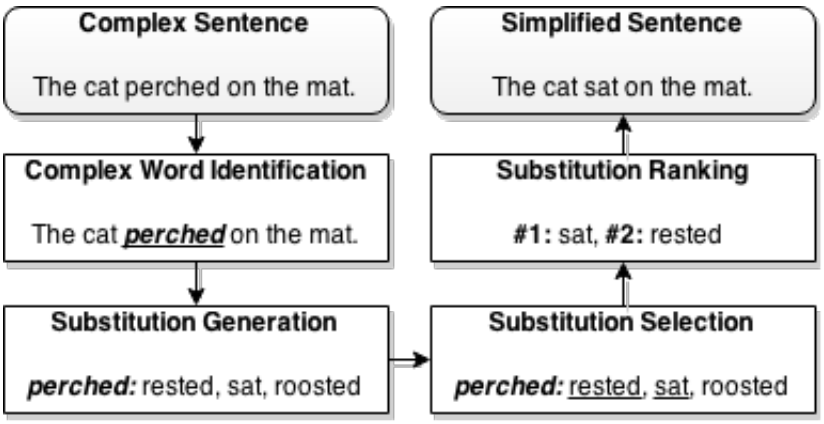

Fig. 1.1 Lexical Simplification pipeline

Each task in Figure 1.1 can be described as:

1. **Complex Word Identification**: Task of deciding which words of a given sentence may not be understood by a given target audience and hence must be simplified.
2. **Substitution Generation**: Task of finding pairs of words or expressions that share the same meaning and are interchangeable in some context.
3. **Substitution Selection**: Task of deciding which is the meaning of a given complex ambiguous word in a sentence to be simplified, and then selecting which substitutions available also represent that meaning.

4. **Substitution Ranking**: Task of ranking the remaining substitutions of a given complex word by their simplicity.

Its name is a reference to the *Frankenstein* novel, published by [33]. It refers to the fact that one can, as Victor Frankenstein did, create an entire Lexical Simplification "creature" by combining multiple "pieces" of the Lexical Simplification pipeline together.

LEXenstein is divided in several modules: Complex Word Identification, Substitution Generation, Substitution Selection, Substitution Ranking, Feature Estimation, Evaluation, Text Adorning, Spelling Correction and Utilities. In the next Sections, we describe the VICTOR and CWICTOR file formats, explain how to properly setup LEXenstein, and discuss in detail the components included in each module of the framework.

Chapter 2

Installation

LEXenstein is a library written entirely in Python. It hence requires for Python 2.7.* to be installed in the user's machine. We have not yet tried to run LEXenstein over Python 3.*. To install LEXenstein, please follow the steps below:

1. Download and unpack the tool from http://ghpaetzold.github.io/LEXenstein/.

2. Navigate to the tool's root folder.

3. Run the following command line:

```
python setup.py install
```

4. If you don't with to install the tool in your Python distribution, you can alternatively copy the "lexenstein" folder into the folder of your project.

5. Access LEXenstein by importing its modules in the following fashion:

```
from lexenstein.morphadorner import *
from lexenstein.spelling import *
from lexenstein.features import *
from lexenstein.identifiers import *
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.rankers import *
from lexenstein.evaluators import *
from lexenstein.util import *
```

LEXenstein requires for several libraries and toolkits to be included in the user's Python 2.7.* installation. The following Sections explain which libraries and toolkits are required, where to get them and how to install them.

2.1 Required Tools and Libraries

2.1.1 Morph Adorner Toolkit

The Morph Adorner Toolkit [25] is a set of Java applications that facilitate the access to Morph Adorner's functionalities [7]. This tool is used by LEXenstein's Substitution Generation module to create inflections for generated substitutions. To install it, follow the steps below:

1. Download the tool from http://ghpaetzold.github.io/MorphAdornerToolkit/
2. Place it in a folder of your choice

Since the tool does not require any compilation, all you need to do is use the path in which you installed it to create instances of the **MorphAdornerToolkit** class, which can be found in LEXenstein's Morph Adorning module.

2.1.2 NLTK

NLTK [3] is a set of resources and algorithms for tasks related, but not restricted to, Natural Language Processing. To install it, please follow the steps provided in: http://www.nltk.org/install.html. Once you have NLTK installed in your Python distribution, please download all additional resources available by following this tutorial: http://www.nltk.org/data.html.

2.1.3 KenLM

KenLM [14] is a tool for fast language model creation and querying. LEXenstein's modules use KenLM to access the data in binary language models for various tasks, such as feature calculation and substitution filtering. To install it, please follow the steps below:

1. Download or clone KenLM from https://github.com/kpu/kenlm
2. Place it in a folder of your choice
3. Navigate to the installation folder in a terminal and run: **python setup.py install**

If no problems occur, KenLM should now be installed in your Python distribution. To verify whether or not the installation was successful, open Python and try importing the library with the following line of code: **import kenlm**. If no errors occur, then the installation was successful.

2.1.4 Scipy and Numpy

Scipy and Numpy [18] are tools that offer great utility for projects and applications in the fields of mathematics, science, and engineering. To install them, please follow the instructions in: http://www.scipy.org/install.html.

2.1.5 Gensim

Gensim [29] is a set of algorithms for unsupervised semantic modeling. LEXenstein uses Gensim to read word vector models. To install it, follow the instructions in: https://radimrehurek.com/gensim/install.html.

2.1.6 PyWSD

PyWSD [35] is a library that offers access to several Word Sense Disambiguation algorithms. LEXenstein's uses this library to filter substitutions. To install it, follow the steps below:

1. Download or clone PyWSD from https://github.com/alvations/pywsd
2. Place it in a folder of your choice
3. Navigate to the installation folder in a terminal and run: **python setup.py install**

If no problems occur, PyWSD should now be installed in your Python distribution. To verify whether or not the installation was successful, open Python and try importing the library with the following line of code: **import pywsd**. If no errors occur, then the installation was successful.

2.1.7 Scikit-Learn

Scikit-Learn [28] is a set of tools for data mining, data analysis and machine learning. LEXenstein uses this library to learn ranking models. To install it, follow the instructions in: http://scikit-learn.org/stable/install.html.

2.1.8 SVM-Rank

SVM-Rank [17] is a tool that allows for one to use Support Vector Machines in ranking setups. LEXenstein uses this library to learn ranking models. To install it, follow the instructions in: http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html.

2.1.9 Stanford Tagger

The Stanford Tagger [20] is a tool that allows for one to annotate sentences with Part-of-Speech (POS) tags. LEXenstein uses this library to find the POS tag of a target word in a sentence. To install it, download the application's latest version from http://nlp.stanford.edu/software/tagger.shtml. Inside the package, you will be able to find the **stanford-tagger.jar** executable and pre-trained tagging models inside the **/models/** folder required by some of LEXenstein's substitution generators.

2.1.10 Other Libraries

LEXenstein also uses various other well-known Python modules, they are: xml, re, urllib2, subprocess, codecs and os.

2.2 Resources

Throughout the following Chapters, you will find usage examples of all classes and functions in LEXenstein. These examples will refer to various sample resources, such as:

- **corpus.txt**: A corpus of text.
- **spelling_model.bin**: A binary spelling model trained with the NorvigCorrector class from the Spelling Correction module.
- **morph**: The Morph Adorner Toolkit [25].
- **lexicon.txt**: A vocabulary.
- **embeddings_model.bin**: A binary word embeddings model trained with word2vec [23].
- **lm.bin**: A binary language model trained with KenLM [14].
- **translation_probs.bin**: A binary translation probabilities file produced with the "addTranslationProbabilitiesFileToShelve" function from the Utilities module.

- **cond_prob.bin**: A binary POS tag conditional probability model trained with the "createConditionalProbabilityModel" function from the Utilities module.
- **pos_model.tagger**: A POS tagging model in the format used by the Stanford Tagger [20].
- **stanford-postagger.jar**: The Stanford Tagger [20].
- **lexmturk.txt**: A sample of the LexMTurk dataset [15] in VICTOR format.
- **train_cwictor_corpus.txt**: A sample of the training set from the Complex Word Identification task of SemEval 2016 [27] in CWICTOR format.
- **test_cwictor_corpus.txt**: A sample of the test set from the Complex Word Identification task of SemEval 2016 [27] in CWICTOR format.
- **tagged_embeddings_model.bin**: A binary word embedding models trained with word2vec [23] over a corpus annotated with generalized POS tags following the convention used by the "getGeneralisedPOS" function from the Utilities model.
- **parallel.txt**: A file containing complex-to-simple POS tagged aligned sentences.
- **alignments.txt**: A file containing word alignments between the sentences in "parallel.txt".
- **stop_words.txt**: A list of stop words.
- **vocab_complex.txt**: A vocabulary extracted from texts of complex nature.
- **vocab_simple.txt**: A vocabulary extracted from texts of simple nature.
- **lm_complex.bin**: A binary language model trained with KenLM [14] over texts of complex nature.
- **lm_simple.bin**: A binary language model trained with KenLM [14] over texts of simple nature.
- **cooc_model.txt**: A word co-occurrence model.
- **clusters.txt**: A list of word clusters produced with the Brown Clustering algorithm [6].
- **ngrams.bin**: A binary n-gram count file produced with the "addNgramCountsFileToShelve" function fro the Utilities module.

- **pos_ngrams.bin**: A POS tagged binary n-gram count file produced with the "addNgram-CountsFileToShelve" function from the Utilities module.
- **dep_models.jar**: A JAR library containing parsing models trained with the Stanford Parser [20].
- **stanford-parser.jar**: The Stanford Parser [20].
- **dep_counts.bin**: A binary file containing dependency link counts produced with the "dependencyParseSentences" and "addNgramCountsFileToShelve" functions from the Utilities file.

You can download a package containing all of these resources from http://www.quest.dcs.shef.ac.uk/lexenstein/LEXenstein_resources.tar.gz. For more instructions on how to create these resources, please refer to LEXenstein's API documentation and the tutorials in Section 6.2.

Chapter 3

The VICTOR and CWICTOR Formats

In order to standardize input and output within the LEXenstein framework, we have conceived the VICTOR and CWICTOR formats: an elegant way of representing data for both the training and testing of Lexical Simplification models and systems. It is a reference to Victor Frankenstein, the main character in the Frankenstein novel [33], and creator of the ever so popular Frankenstein's Monster.

3.1 The VICTOR Format

The VICTOR format was conceived to represent datasets for the tasks of Substitution Generation, Selection and Ranking. Each line of a file in VICTOR format is structured as illustrated in Example 3.1, where S_i is the ith sentence in the dataset, w_i a target complex word in the h_ith position of S_i, c_i^j a substitution candidate and r_i^j its simplicity ranking.

$$\langle \mathrm{S}_i \rangle \ \langle \mathrm{w}_i \rangle \ \langle \mathrm{h}_i \rangle \ \left\langle \mathrm{r}_i^1 : \mathrm{c}_i^1 \right\rangle \left\langle \mathrm{r}_i^2 : \mathrm{c}_i^2 \right\rangle \cdots \left\langle \mathrm{r}_i^{n-1} : \mathrm{c}_i^{n-1} \right\rangle, \langle \mathrm{r}_i^n : \mathrm{c}_i^n \rangle \tag{3.1}$$

Each bracketed component in the example above is separated by a tabulation marker. Examples of files with such notation can be found in "resources/victor_datasets".

3.2 The CWICTOR Format

The CWICTOR format was conceived to represent datasets for the task of Complex Word Identification. Each line of a file in CWICTOR format is structured as illustrated in Exam-

ple 3.2, where S_i is the ith sentence in the dataset, w_i a the word in the h_ith position of S_i, and l_i a binary label, which must have value 1 if w_i is complex, and value 0 otherwise.

$$\langle S_i \rangle \ \langle w_i \rangle \ \langle h_i \rangle \ \langle l_i \rangle \tag{3.2}$$

Each bracketed component in the example above is separated by a tabulation marker. Examples of files with such notation can be found in “resources/cwictor_datasets”.

Chapter 4

The Spelling Correction Module

LEXenstein's Spelling Correction module (lexenstein.spelling) allows for one to correct misspelled words. The module is used by all generators in order to ensure that words and lemmas don't have their grammaticality compromised during inflection. It includes the NorvigCorrector class, which employs the spelling correction algorithm of Norvig[1].

```
from lexenstein.spelling import *

nc = NorvigCorrector('corpus.txt', format='text')
nc.saveBinaryModel('spelling_model.bin')
nc = NorvigCorrector('spelling_model.bin', format='bin')
print(nc.correct('mystake'))
print(nc.correct('speling'))
print(nc.correct('beauriful'))
```

The output produced by the script above will be:

```
mistake
spelling
beautiful
```

[1] http://norvig.com/spell-correct.html

Chapter 5

The Text Adorning Module

LEXenstein's Text Adorning module (lexenstein.morphadorner) provides a Python interface to the Morph Adorner Toolkit [25], a set of Java tools that facilitates the access to Morph Adorner's functionalities. The class MorphAdornerToolkit provides easy access to word lemmatization, word stemming, syllable splitting, noun inflection, verb tensing, verb conjugation and adjective/adverb inflection. Below is an example of how to create and use a MorphAdornerToolkit object:

```
from lexenstein.morphadorner import MorphAdornerToolkit

m = MorphAdornerToolkit('./morph/')

lemmas = m.lemmatizeWords(['doing', 'geese'])
print('Lemmas:')
print(str(lemmas)+'\n')

stems = m.stemWords(['doing', 'geese'])
print('Stems:')
print(str(stems)+'\n')

tenses = m.tenseVerbs(['do'], ['doing'])
print('Tenses:')
print(str(tenses)+'\n')

verbs = m.conjugateVerbs(['do', 'sit'], 'PRESENT_PARTICIPLE',
    'THIRD_PERSON_PLURAL')
print('Verbs:')
print(str(verbs)+'\n')
```

```
nouns = m.inflectNouns(['goose', 'chair'], 'plural')
print('Nouns:')
print(str(nouns)+'\n')

syllables = m.splitSyllables(['persevere', 'sitting'])
print('Syllables:')
print(str(syllables)+'\n')

adjectives = m.inflectAdjectives(['nice', 'pretty'], 'comparative')
print('Adjectives:')
print(str(adjectives)+'\n')
```

The output produced by the script above will be:

```
Lemmas:
['do', 'goose']

Stems:
['do', 'gees']

Tenses:
[['PRESENT_PARTICIPLE', 'THIRD_PERSON_PLURAL']]

Verbs:
['doing', 'sitting']

Nouns:
['geese', 'chairs']

Syllables:
['per-se-vere', 'sit-ting']

Adjectives:
['nicer', 'prettier']
```

Chapter 6

The Feature Estimation Module

LEXenstein's Feature Estimation module (lexenstein.features) allows the calculation of several features for LS related tasks. Its class FeatureEstimator allows the user to select and configure many types of features commonly used by LS approaches.

The FeatureEstimator object can be used either for the creation of LEXenstein's rankers, or in stand-alone setups. For the latter, the class provides a function called *calculateFeatures*, which takes as input a dataset in the VICTOR/CWICTOR format (that can be built from generated/selected substitutions). If the dataset is in VICTOR format, it returns as output a matrix *M*x*N* containing *M* feature values for each of the *N* substitution candidates listed in the dataset. If the dataset is in CWICTOR format, it returns as output a matrix *M*x*N* containing *M* feature values for the target word of each of the *N* instances of the dataset.

6.1 Features Supported

Each of the 39 features supported must be configured individually. They can be grouped in eight categories:

- **Lexicon-oriented**: A binary feature that receives value 1 if a candidate appears in a given vocabulary, and 0 otherwise.
- **Morphological**: Features that exploit morphological characteristics of substitutions. They include:

    1. The length of a candidate.
    2. The number of syllables of a candidate.

- **Collocational**: Comprised of several raw frequency counts and language model probabilities of the form $P\left(S_{h-1}^{h-l}\, c\, S_{h+1}^{h+r}\right)$, where c is a candidate substitution in the hth position in sentence S, and S_{h-1}^{h-l} and S_{h+1}^{h+r} are n-grams of size l and r, respectively. Included in this category of features are:

  1. The language model probability of the entire sentence S with the target word replace by a candidate.
  2. The language model probability of an n-gram.
  3. The raw frequency count of an n-gram in a corpus.
  4. A binary feature that receives value 1 if an n-gram is in a corpus, and 0 otherwise.

- **Pop-Collocational**: Comprised of several raw frequency counts and language model probabilities of the "pop" n-grams introduced by [16]. They differ from typical collocational features in the sense that, instead of retrieving the frequency of the n-gram itself, it retrieves the highest frequency between three variants of an n-gram: its original form, $S_{h-1}^{h-l}\, c\, S_{h+1}^{h+r}$, its leftmost "popped" version, $S_{h-2}^{h-l}\, c\, S_{h+1}^{h+r}$, its rightmost "popped" version, $S_{h-1}^{h-l}\, c\, S_{h+2}^{h+r}$, and its double "popped" version, $S_{h-2}^{h-l}\, c\, S_{h+2}^{h+r}$. Included in this category of features are:

  1. The language model probability of a pop n-gram.
  2. The raw frequency count of a pop n-gram in a corpus.

- **Tagged-Collocational**: Comprised of several raw frequency counts of "tagged" n-grams, a novel concept introduced in this work. They differ from typical collocational features in the sense that, instead of retrieving the frequency of an n-gram in its original form, $S_{h-1}^{h-l}\, c\, S_{h+1}^{h+r}$, it retrieves the frequency of a candidate surrounded by the neighbor words' POS tags, $P_{h-1}^{h-l}\, c\, P_{h+1}^{h+r}$, given that P is the set of POS tags that describe sentence S. Included in this category of features are:

  1. The raw frequency count of a tagged n-gram in a corpus.
  2. A binary feature that receives value 1 if a tagged n-gram is in a corpus, and 0 otherwise.

- **Sense-oriented**: Comprised of features that describe the semantic information of a candidate substitution. They include:

  1. The number of senses of a candidate.

2. The number of synonyms of a candidate.
3. The number of hypernyms of a candidate.
4. The number of hyponyms of a candidate.
5. The maximum semantic distance between all of a candidate's senses in a thesaurus.
6. The minimum semantic distance between all of a candidate's senses in a thesaurus.
7. A binary feature that receives value 1 if the candidate is a synonym of the target word, and 0 otherwise.
8. A binary feature that receives value 1 if the candidate is a hypernym of the target word, and 0 otherwise.
9. A binary feature that receives value 1 if the candidate is a hyponym of the target word, and 0 otherwise.

These feature values are extracted from the WordNet thesaurus.

- **Syntactic**: Comprised of features that measure how likely a candidate substitution is of assuming the syntactic role of the target word. We describe the syntactic role of a target word w in sentence S as its POS tag and its set of subject dependency links, $(w \rightarrow S_i)$, and object dependency links, $(w \leftarrow S_i)$. Included in this category of features are:

  1. The conditional probability of the candidate assuming the POS tag of the target word.
  2. The average probability of the candidate assuming the subject dependency links of the target word.
  3. The average raw occurrence counts in which the candidate assumes the subject dependency links of the target word.
  4. A binary feature that receives value 1 if the candidate assumes all subject dependency links of the target word at least once in a corpus.
  5. The average probability of the candidate assuming the object dependency links of the target word.
  6. The average raw occurrence counts in which the candidate assumes the object dependency links of the target word.

7. A binary feature that receives value 1 if the candidate assumes all object dependency links of the target word at least once in a corpus.
8. The average probability of the candidate assuming all dependency links of the target word.
9. The average raw occurrence counts in which the candidate assumes all dependency links of the target word.
10. A binary feature that receives value 1 if the candidate assumes all dependency links of the target word at least once in a corpus.

- **Semantic**: Comprised of features that describe not only the semantic content pertaining to a candidate individually, but also its semantic similarity with the target word and the sentence in question. Included in this category of features are:

1. The probability of the target word being translated into a given candidate.
2. The candidate's word embedding values, as determined by a word embeddings model.
3. The cosine similarity between the candidate's and the target word's embedding vectors.
4. The average cosine similarity between the candidate's embeddings vector, and those of all content words in the sentence.
5. The candidate's word embedding values, as determined by an enhanced word embeddings model.
6. The cosine similarity between the candidate's and the target word's enhanced embedding vectors.
7. The average cosine similarity between the candidate's enhanced embeddings vector, and those of all content words in the sentence.

The code snippet below shows an example of how to configure and use a FeatureEstimator object:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.features import *

m = MorphAdornerToolkit('./morph/')

#Numerical features:
```

```
fe = FeatureEstimator(norm=False)
fe.addLexiconFeature('lexicon.txt', 'Simplicity')
fe.addLengthFeature('Complexity')
fe.addSyllableFeature(m, 'Complexity')
fe.addCollocationalFeature('lm.bin', 2, 2, 'Simplicity')
fe.addFrequencyCollocationalFeature('ngrams.bin', 2, 2, 'Simplicity')
fe.addTaggedFrequencyCollocationalFeature('pos_ngrams.bin', 2, 2,
    './pos_model.tagger', './stanford-postagger.jar', '/usr/bin/java',
    pos_tag='treebank', 'Simplicity')
fe.addBinaryTaggedFrequencyCollocationalFeature('pos_ngrams.bin', 2, 2,
    './pos_model.tagger', './stanford-postagger.jar', '/usr/bin/java',
    pos_tag='treebank', 'Simplicity')
fe.addPopCollocationalFeature('lm.bin', 2, 2, 'Simplicity')
fe.addNGramProbabilityFeature('lm.bin', 2, 2, 'Simplicity')
fe.addNGramFrequencyFeature('ngrams.bin', 2, 2, 'Simplicity')
fe.addBinaryNGramFrequencyFeature('ngrams.bin', 2, 2, 'Simplicity')
fe.addPopNGramProbabilityFeature('lm.bin', 2, 2, 'Simplicity')
fe.addPopNGramFrequencyFeature('ngrams.bin', 2, 2, 'Simplicity')
fe.addNGramFrequencyFeature('lm.bin', 3, 0, 'Simplicity')
fe.addSentenceProbabilityFeature('lm.bin', 'Simplicity')
fe.addSenseCountFeature('Simplicity')
fe.addSynonymCountFeature('Simplicity')
fe.addIsSynonymFeature('Simplicity')
fe.addHypernymCountFeature('Simplicity')
fe.addIsHypernymFeature('Simplicity')
fe.addHyponymCountFeature('Simplicity')
fe.addIsHyponymFeature('Simplicity')
fe.addMinDepthFeature('Complexity')
fe.addMaxDepthFeature('Complexity')
fe.addAverageDepthFeature('Complexity')
fe.addTranslationProbabilityFeature('translation_probs.bin', 'Simplicity')
fe.addWordVectorValues('embeddings_model.bin', 100, 'Simplicity')
fe.addWordVectorSimilarityFeature('embeddings_model.bin', 'Simplicity')
fe.addTaggedWordVectorSimilarityFeature('tagged_embeddings_model.bin',
    './pos_model.tagger', './stanford-postagger.jar', '/usr/bin/java',
    'paetzold', 'Simplicity')
fe.addWordVectorContextSimilarityFeature('embeddings_model.bin',
    './pos_model.tagger', './stanford-postagger.jar', '/usr/bin/java',
    'Simplicity')
```

```
fe.addTaggedWordVectorContextSimilarityFeature('tagged_embeddings_model.bin',
    './pos_model.tagger', './stanford-postagger.jar', '/usr/bin/java',
    'paetzold', 'Simplicity')
fe.addTargetPOSTagProbability('cond_prob.bin', './pos_model.tagger',
    './stanford-postagger.jar', '/usr/bin/java', 'Simplicity')
fe.addSubjectDependencyProbabilityFeature('dep_lm.bin',
    './stanford-parser.jar', './dep_models.jar', '/usr/bin/java',
    'Simplicity')
fe.addObjectDependencyProbabilityFeature('dep_lm.bin',
    './stanford-parser.jar', './dep_models.jar', '/usr/bin/java',
    'Simplicity')
fe.addAllDependencyProbabilityFeature('dep_lm.bin', './stanford-parser.jar',
    './dep_models.jar', '/usr/bin/java', 'Simplicity')
fe.addBinarySubjectDependencyFeature('dep_counts.bin',
    './stanford-parser.jar', './dep_models.jar', '/usr/bin/java',
    'Simplicity')
fe.addBinaryObjectDependencyFeature('dep_counts.bin',
    './stanford-parser.jar', './dep_models.jar', '/usr/bin/java',
    'Simplicity')
fe.addBinaryAllDependencyFeature('dep_counts.bin', './stanford-parser.jar',
    './dep_models.jar', '/usr/bin/java', 'Simplicity')
fe.addSubjectDependencyFrequencyFeature('dep_counts.bin',
    './stanford-parser.jar', './dep_models.jar', '/usr/bin/java',
    'Simplicity')
fe.addObjectDependencyFrequencyFeature('dep_counts.bin',
    './stanford-parser.jar', './dep_models.jar', '/usr/bin/java',
    'Simplicity')
fe.addAllDependencyFrequencyFeature('dep_counts.bin',
    './stanford-parser.jar', './dep_models.jar', '/usr/bin/java',
    'Simplicity')
fe.addBackoffBehaviorNominalFeature('ngrams.bin', 'Simplicity')
fe.addImageSearchCountFeature(00000, 'Simplicity')
fe.addMorphologicalFeature(my_dict, 'my_feature', 'Simplicity')

#Nominal features:
fe.addNullLinkNominalFeature('./stanford-parser.jar', './dep_models.jar',
    '/usr/bin/java', 'Simplicity')
fe.addCandidateNominalFeature()
fe.addNgramNominalFeature(2, 2)
```

```
fe.addCandidatePOSNominalFeature('./pos_model.tagger',
    './stanford-postagger.jar', '/usr/bin/java', pos_type='treebank')
fe.addPOSNgramNominalFeature(2, 2, './pos_model.tagger',
    './stanford-postagger.jar', '/usr/bin/java', pos_type='treebank')
fe.addPOSNgramWithCandidateNominalFeature(2, 2, './pos_model.tagger',
    './stanford-postagger.jar', '/usr/bin/java', pos_type='treebank')

feats = fe.calculateFeatures('lexmturk.txt', format='victor')
```

The output produced by the script above will be:

```
[[1.0, 7, 2, 4.470454216003418, 6.206913948059082, 7.3130669593811035,
    10.297701835632324, 12.034161567687988, 13.140314102172852,
    11.422476768493652, 13.158936500549316, 14.26508903503418,
    51.03892517089844, 12, 51, 12, 83, 3, 8, 0.0167457456, 0.234516], [0.0,
    6, 3, 4.487872123718262, 7.389286041259766, 8.941888809204102,
    10.315119743347168, 13.216533660888672, 14.769136428833008,
    11.439894676208496, 14.34130859375, 15.893911361694336, 52.66774368286,
    2, 6, 0, 0, 0, 0, 0.0134387482374, 0.234833]
...
[0.0, 7, 2, 4.950876235961914, 8.496809005737305, 9.830375671386719,
    10.77812385559082, 14.324056625366211, 15.657623291015625,
    11.902898788452148, 15.448831558227539, 16.782398223876953,
    53.556236267089844, 6, 13, 1, 0, 0, 8, 0.0023717232, 0.43812894]]
```

In the Section that follows, we elaborate on how to produce the resources required by the features included in LEXenstein.

6.2 Producing Feature Resources

In this Section, we provide tutorials on how to produce several resources required by LEXenstein's features. They are:

- **Lexicons**: Must be a plain text file with one word per line.
- **Language Models**: Must be produced by KenLM with the following command lines:

```
lmplz -o [order] <[corpus_of_text] >[language_model]
```

```
build_binary [language_model] [binary_language_model]
```

The user can also use other language modeling tools, such as SRILM, to produce a language model in ARPA format, and then binarize it with KenLM.

- **N-gram Count Files**: Must be a Python shelve file created with the "addNgramCountsFileToShelve" function in the Utilities module of LEXenstein (lexenstein.util). For more detailed instructions, please refer to the documentation of the "addNgramCountsFileToShelve" function.

- **Tagged N-gram Count Files**: Must be a Python shelve file created with the "addNgramCountsFileToShelve" function in the Utilities module of LEXenstein (lexenstein.util). To create a tagged n-grams file using SRILM, use the "createTaggedNgramsFile" function in the Utilities module. For more detailed instructions, please refer to the documentation of the "addNgramCountsFileToShelve" and "createTaggedNgramsFile" functions.

- **Dependency Link Language Models**: To create such a language model, you must:

  1. Produce a large corpus of dependency parsed sentences.
  2. Transform each and every dependency link to the format of Example 6.1, where each token is space-separated.

  $$<\text{type\_of\_dependency\_link}><\text{subject\_word}><\text{object\_word}> \quad (6.1)$$

  3. Place all dependency links in a plain text file.
  4. Run a language modeling tool over the file, producing a file in ARPA format.
  5. Binarize the language model with KenLM using the following command:

  ```
  build_binary [language_model] [binary_language_model]
  ```

- **Translation Probability Files**: Must be a Python shelve file created with the "addTranslationProbabilitiesFileToShelve" function in the Utilities module of LEXenstein (lexenstein.util). For more detailed instructions, please refer to the documentation of the "addTranslationProbabilitiesFileToShelve" function.

- **Word Embedding Models**: Must be a binary file created through the use of word2vec. For more information on how to create them, please follow the instructions on the website of the application at https://code.google.com/p/word2vec/.

- **Tagged Word Vector Models**: To create a tagged word vector model, you must:

  1. Produce POS tags for a large corpus of text.
  2. Concatenate the POS tags to each word in the corpus using the format of Example 8.1, where w_i is the ith word in a sentence, and p_i its POS tag.

  $$w_1|||p_1\ w_2|||p_2\ ...\ w_{n-1}|||p_{n-1}\ w_n|||p_n \tag{6.2}$$

  3. Train a binary word vector model over the resulting corpus using word2vec.

  LEXenstein supports two POS tag conventions:

  1. Treebank: POS tags in the Penn Treebank format [22]. They can be produced by any modern POS tagger, such as the Stanford Tagger [20].
  2. Paetzold: Generalized versions of Treebank tags. They can be derived from Treebank tags using the "getGeneralisedPOS" from the LEXenstein's Utilities module (lexenstein.util).

- **Conditional Probability Models**: Must be created by the "createConditionalProbabilityModel" function from LEXenstein's Utilities module (lexenstein.utilities). For detailed instructions, please refer to the documentation of the "createConditionalProbabilityModel" function.

Chapter 7

The Complex Word Identification Module

We define Complex Word Identification (CWI) as the task of deciding which words can be considered complex by a certain target audience. There are not many examples in literature of approaches for the task. In [31] the performance of three approaches are compared over the dataset introduced by [32]. The LS strategy of [15] provides an implicit approach for the task.

In the Complex Word Identification module of LEXenstein (lexenstein.identifiers), one has access to several Complex Word Identification methods. The module contains a series of classes, each representing a distinct method. Currently, it offers support for 5 distinct approaches. The following Sections describe each one individually.

7.1 MachineLearningIdentifier

Uses Machine Learning techniques to train classifiers that distinguish between complex and simple words. The class allows for the user to train models with 4 Machine Learning techniques: Support Vector Machines, Decision Trees and linear models estimated with Stochastic Gradient Descent and Passive Aggressive Learning. As input, it requires for a FeatureEstimator object. As output, it produces a vector containing a binary label for the target word of each of the N instances in a dataset in VICTOR or CWICTOR format. The label will have value 1 if the target word was predicted as complex, and 0 otherwise.

7.1.1 Parameters

During instantiation, the MachineLearningIdentifier requires only for a FeatureEstimator object. The user can then use the "calculateTrainingFeatures" and "calculateTestingFeatures" functions to estimate features of training and testing datasets in VICTOR or CWICTOR formats, the "selectKBestFeatures" function to apply feature selection over the features estimated, the "trainSVM", "trainDecisionTreeClassifier", "trainSGDClassifier" and "trainPassiveAggressiveClassifier" functions to train CWI models, and finally the "identifyComplexWords" function to produce output labels. To know more about the parameters and recommended values of the aforementioned functions, please refer to LEXenstein's documentation.

7.1.2 Example

The code snippet below shows the MachineLearningIdentifier class being used:

```
from lexenstein.identifiers import *
from lexenstein.features import *

fe = FeatureEstimator()
fe.addLexiconFeature('lexicon.txt', 'Simplicity')
fe.addLengthFeature('Complexity')
fe.addSenseCountFeature('Simplicity')

mli = MachineLearningIdentifier(fe)
mli.calculateTrainingFeatures('train_cwictor_corpus.txt')
mli.calculateTestingFeatures('test_cwictor_corpus.txt')
mli.selectKBestFeatures(k=2)
mli.trainDecisionTreeClassifier()

labels = mli.identifyComplexWords()
```

7.2 LexiconIdentifier

Uses a lexicon of complex/simple words in order to judge the complexity of a word: if it appears in the lexicon, then it is complex/simple. As input, it requires for a lexicon file of complex or simple words and expressions. As output, it produces a vector containing a binary label for the target word of each of the N instances in a dataset in VICTOR or CWICTOR

format. The label will have value 1 if the lexicon classifies the word as complex, and 0 otherwise.

7.2.1 Parameters

During instantiation, the LexiconIdentifier requires for a lexicon file and an indicator label. The lexicon file must be in plain text and contain one word/expression per line, and the value of the label must be either "complex" or "simple". If the label's value is "complex", then the lexicon will be interpreted as a vocabulary of complex words and expressions, otherwise it will be interpreted as a vocabulary of simple words.

7.2.2 Example

The code snippet below shows the LexiconIdentifier class being used:

```
from lexenstein.identifiers import *

li = LexiconIdentifier('lexicon.txt', 'simple')

labels = li.identifyComplexWords('test_cwictor_corpus.txt')
```

7.3 ThresholdIdentifier

Estimates the threshold t over a given feature value that best separates complex and simple words. As input, it requires for a FeatureEstimator object. As output, it produces a vector containing a binary label for the target word of each of the N instances in a dataset in VICTOR or CWICTOR format. The label will have value 1 if the lexicon classifies the word as complex, and 0 otherwise.

7.3.1 Parameters

During instantiation, the ThresholdIdentifier requires for a FeatureEstimator object. The user can then use the "calculateTrainingFeatures" and "calculateTestingFeatures" functions to estimate features of training and testing datasets in VICTOR or CWICTOR formats, the "trainIdentifierBruteForce" and "trainIdentifierBinarySearch" functions to train CWI models, and finally the "identifyComplexWords" function to produce output labels.

The “trainIdentifierBruteForce” and “trainIdentifierBinarySearch” functions require for a feature index as input. It will determine over which feature of the ones included in the FeatureEstimator object the threshold t will be estimated. To know more about the parameters and recommended values of the aforementioned functions, please refer to LEXenstein’s documentation.

7.3.2 Example

The code snippet below shows the ThresholdIdentifier class being used:

```
from lexenstein.identifiers import *
from lexenstein.features import *

fe = FeatureEstimator()
fe.addLengthFeature('Complexity')
fe.addSenseCountFeature('Simplicity')

ti = ThresholdIdentifier(fe)
ti.calculateTrainingFeatures('train_cwictor_corpus.txt')
ti.calculateTestingFeatures('test_cwictor_corpus.txt')
ti.trainIdentifierBruteForce(1)

labels = ti.identifyComplexWords()
```

Chapter 8

The Substitution Generation Module

We define Substitution Generation (SG) as the task of producing candidate substitutions for complex words. Authors commonly address this task by querying WordNet [12] and UMLS[4]. Some examples of authors who resort to this strategy are [11] and [9]. Recently however, learning substitutions from aligned corpora have become a more popular strategy [26] and [15].

In the Substitution Generation module of LEXenstein (lexenstein.generators), one has access to several Substitution Generation methods available in literature. The module contains a series of classes, each representing a distinct approach in literature. Currently, it offers support for 5 distinct approaches. All approaches use LEXenstein's Text Adorning module, described in Section 5, to create substitutions for all possible inflections of verbs and nouns. The following Sections describe each one individually.

8.1 PaetzoldGenerator

Employs a novel strategy, in which substitutions are extracted from tagged word embedding models. To be instantiated, this class requires as input a path to a binary tagged word vector model trained with word2vec. As output, its *getSubstitutions* function produces a dictionary containing the *n* words of which the embeddings vector has the highest cosine similarity with each target word in a VICTOR corpus.

8.1.1 Parameters

The word vector model required by the PaetzoldGenerator class must be in the binary format produced by word2vec. The model can be trained over a POS tag annotated corpus using any of the parameters supported by word2vec. To produce the tagged corpus required, you must:

1. Produce POS tags for a large corpus of text.
2. Concatenate the POS tags to each word in the corpus using the format of Example 8.1, where w_i is the ith word in a sentence, and p_i its POS tag.

$$w_1|||p_1\ w_2|||p_2\ \ldots\ w_{n-1}|||p_{n-1}\ w_n|||p_n \tag{8.1}$$

3. Train a binary word vector model over the resulting corpus using word2vec.

The PaetzoldGenerator supports two POS tag conventions:

1. Treebank: POS tags in the Penn Treebank format [22]. They can be produced by any modern POS tagger, such as the Stanford Tagger [20].
2. Paetzold: Generalized versions of Treebank tags. They can be derived from Treebank tags using the "getGeneralisedPOS" from the LEXenstein's Utilities module (lexenstein.util).

To learn how to use word2vec, please refer to their documentation at https://code.google.com/p/word2vec/.

8.1.2 Example

The code snippet below shows the PaetzoldGenerator class being used:

```
from lexenstein.generators import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

pg = PaetzoldGenerator('tagged_embeddings_model.bin', nc,
    'pos_model.tagger', 'stanford-postagger.jar', '/usr/bin/java')
subs = pg.getSubstitutions('lexmturk.txt', 10)
```

8.2 GlavasGenerator

Employs the strategy described in [13], in which substitutions are extracted from typical word embedding models. To be instantiated, this class requires as input a path to a binary word vector model trained with word2vec. As output, its *getSubstitutions* function produces

a dictionary containing the n words of which the embeddings vector has the highest cosine similarity with each target word in a VICTOR corpus.

8.2.1 Parameters

The word vector model required by the GlavasGenerator class must be in the binary format produced by word2vec. The model can be trained over any type of corpus using any of the parameters supported by word2vec. To learn how to use word2vec, please refer to their documentation at https://code.google.com/p/word2vec/.

8.2.2 Example

The code snippet below shows the GlavasGenerator class being used:

```
from lexenstein.generators import *

kg = GlavasGenerator('embeddings_model.bin')
subs = kg.getSubstitutions('lexmturk.txt', 10)
```

8.3 KauchakGenerator

Employs the strategy described in [15], in which substitutions are automatically extracted from parallel corpora. To be instantiated, this class requires as input an object of the MorphAdornerToolkit class, a parsed document of parallel sentences, the word alignments between them in Pharaoh format, a list of stop words and a NorvigCorrector object. As output, its *getSubstitutions* function produces a dictionary of complex-to-simple substitutions filtered by the criteria described in [15].

8.3.1 Parameters

The parsed parallel document, the alignments file, and the stop words list required by the KauchakGenerator class must be in a specific format. Each line of the parsed parallel document must be in the format described in Example 8.2, where w_i^s is a word in position i of a source sentence s, p_i^s its POS tag, w_j^t is a word in position j of a target sentence t, and p_j^t its POS tag.

$$\langle w_0^s\rangle|||\langle p_0^s\rangle\cdots\langle w_n^s\rangle|||\langle p_n^s\rangle \quad \langle w_0^t\rangle|||\langle p_0^t\rangle\cdots\langle w_n^t\rangle|||\langle p_n^t\rangle \tag{8.2}$$

All tokens of form $\langle w_i^s \rangle ||| \langle p_i^s \rangle$ are separated by a blank space, and the two set of source and target tokens $\langle w_0^s \rangle ||| \langle p_0^s \rangle \cdots \langle w_n^s \rangle ||| \langle p_n^s \rangle$ are separated by a tabulation marker. An example of file with such notation can be found in "resources/parallel_data/alignment_pos_file.txt".

The alignments file must be in Pharaoh format. Each line of the alignments file must be structured as illustrated in Example 8.3, where $\langle i_h^s \rangle$ is an index i in source sentence s, and $\langle j_h^t \rangle$ is the index j in source sentence t aligned to it.

$$\langle i_0^s \rangle - \langle j_0^t \rangle \ \langle i_1^s \rangle - \langle j_1^t \rangle \cdots \langle i_{n-1}^s \rangle - \langle j_{n-1}^t \rangle \ \langle i_n^s \rangle - \langle j_n^t \rangle \tag{8.3}$$

All tokens of form $\langle i_h^s \rangle - \langle j_h^t \rangle$ are separated by a blank space. An example of file with such notation can be found in "resources/parallel_data/alignments.txt".

8.3.2 Example

The code snippet below shows the KauchakGenerator class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

kg = KauchakGenerator(m, 'parallel.txt', 'alignments.txt', 'stop_words.txt',
    nc)
subs = kg.getSubstitutions('lexmturk.txt')
```

8.4 YamamotoGenerator

Employs the strategy described in [19], in which substitutions are extracted from dictionary definitions of complex words. This approach requires as input an API key for the Merriam Dictionary[1], which can be obtained for free, and a NorvigCorrector object. As output, it produces a dictionary linking words in the Merriam Dictionary and WordNet to words with the same Part-of-Speech (POS) tag in its entries' definitions and usage examples.

[1] http://www.dictionaryapi.com/

8.4.1 Parameters

The YamamotoGenerator class requires a free Dictionary key to the Merriam Dictionary. To get the key, follow the steps below:

1. Visit the page http://www.dictionaryapi.com/register/index.htm.
2. Fill in your personal information.
3. In "Request API Key #1:" and "Request API Key #2:", select "Collegiate Dictionary" and "Collegiate Thesaurus".
4. Login in http://www.dictionaryapi.com.
5. Visit your "My Keys" page.
6. Use the "Dictionary" key to create a YamamotoGenerator object.

8.4.2 Example

The code snippet below shows the YamamotoGenerator class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

yg = YamamotoGenerator(m, '0000-0000-0000-0000', nc)
subs = yg.getSubstitutions('lexmturk.txt')
```

8.5 MerriamGenerator

Extracts a dictionary linking words to their synonyms, as listed in the Merriam Thesaurus. This approach requires as input an API key for the Merriam Thesaurus[2], which can be obtained for free, and a NorvigCorrector object.

[2]http://www.dictionaryapi.com/

8.5.1 Parameters

The MerriamGenerator class requires a free Thesaurus key to the Merriam Dictionary. To get the key, follow the steps below:

1. Visit the page http://www.dictionaryapi.com/register/index.htm.
2. Fill in your personal information.
3. In "Request API Key #1:" and "Request API Key #2:", select "Collegiate Dictionary" and "Collegiate Thesaurus".
4. Login in http://www.dictionaryapi.com.
5. Visit your "My Keys" page.
6. Use the "Thesaurus" key to create a MerriamGenerator object.

8.5.2 Example

The code snippet below shows the MerriamGenerator class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

mg = MerriamGenerator(m, '0000-0000-0000-0000', nc)
subs = mg.getSubstitutions('lexmturk.txt')
```

8.6 WordnetGenerator

Extracts a dictionary linking words to their synonyms, as listed in WordNet. It requires for a NorvigCorrector object, the path to a POS tagging model, and the path to the Stanford Tagger.

8.6.1 Parameters

In order to obtain the model and tagger required by the WordnetGenerator class, download the full version of the Stanford Tagger package from the link: http://nlp.stanford.edu/software/tagger.shtml. Inside the package's "models" folder you will find tagging models for various languages. In the package's root folder, you will find the "stanford-postagger.jar" application, which is the one required by the WordnetGenerator.

8.6.2 Example

The code snippet below shows the WordnetGenerator class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')
```

8.7 BiranGenerator

Employs the strategy described in [2], in which substitutions are filtered from the Cartesian product between vocabularies of complex and simple words. This approach requires as input vocabularies of complex and simple words, as well as two Language Models trained over complex and simple corpora, a NorvigCorrector object, the path to a POS tagging model, and the path to the Stanford Tagger. As output, it produces a dictionary linking words to a set of synonyms and hypernyms filtered by the criteria described in [2].

8.7.1 Parameters

The vocabularies of complex and simple words and Language Models trained over complex and simple corpora required by the BiranGenerator class must be in a specific format. The

vocabularies must contain one word per line, and can be produced over large corpora with SRILM by running the following command line:

```
ngram-count -text [corpus_of_text] -write-vocab [vocabulary_name]
```

The Language Models must be binary, and must be produced by KenLM with the following command lines:

```
lmplz -o [order] <[corpus_of_text] >[language_model_name]
```

```
build_binary [language_model_name] [binary_language_model_name]
```

Complex and simple data can be downloaded fro David Kauchak's page[3], or extracted from other sources. In order to obtain the model and tagger required by the BiranGenerator class, download the full version of the Stanford Tagger package from the link: http://nlp.stanford.edu/software/tagger.shtml. Inside the package's "models" folder you will find tagging models for various languages. In the package's root folder, you will find the "stanford-postagger.jar" application, which is the one required by the BiranGenerator.

8.7.2 Example

The code snippet below shows the BiranGenerator class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

bg = BiranGenerator(m, 'vocab_complex.txt', 'vocab_simple.txt',
    'lm_complex.bin', 'lm_simple.bin', nc, 'pos_model.tagger',
    'stanford-postagger.jar', '/usr/bin/java')
subs = bg.getSubstitutions('lexmturk.txt')
```

[3]http://www.cs.pomona.edu/ dkauchak/simplification/

Chapter 9

The Substitution Selection Module

Substitution Selection (SS) is the task of selecting which substitutions from a given list can replace a complex word in a given sentence without altering its meaning. Most work addresses this task referring to the context of the complex word by employing Word Sense Disambiguation (WSD) approaches [24, 30], or by discarding substitutions which do not share the same POS tag of the target complex word [19, 26].

LEXenstein's SS module (lexenstein.selectors) provides access to 8 approaches, each represented by a Python class. All classes have a "selectCandidates" function, which receives as input a set of candidate substitutions and a corpus in the VICTOR format. The candidate substitutions can be either a dictionary produced by a Substitution Generation approach, or a list of candidates that have been already selected by another Substitution Selection approach. This feature allows for multiple selectors to be used in sequence. The following Sections describe each of the classes in the LEXenstein SS module individually.

9.1 WordVectorSelector

Employs a novel strategy, in which a word vector model is used to determine which substitutions have the closest meaning to that of the sentence being simplified. It retrieves a user-defined percentage of the substitutions, which are ranked with respect to the cosine distance between its word vector and the sum of some, or all of the sentences' words, depending on the settings defined by the user.

9.1.1 Parameters

To create a WordVectorSelector object, you must provide a binary word vector model created with Word2Vec. To create the word vector, follow the steps below:

1. Download and install Word2Vec in your machine from https://code.google.com/p/word2vec/.
2. Gather large amounts of corpora (>10 billion words). You can find some sources of data in https://code.google.com/p/word2vec/.
3. With Word2Vec installed, run it with the following command line:

```
./word2vec -train <corpus> -output <binary_model_path> -cbow 1 -size
    300 -window 5 -negative 3 -hs 0 -sample 1e-5 -threads 12 -binary 1
    -min-count 10
```

The command line above creates word vectors with 300 dimensions and considers only a window of 5 tokens around each word. You can customize those parameters if you wish. For more information on how to use other class functions, please refer to LEXenstein's documentation.

9.1.2 Example

The code snippet below shows the WordVectorSelector class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

wordvecselector = WordVectorSelector('embeddings_model.bin',
    'pos_model.tagger', 'stanford-postagger.jar', '/usr/bin/java')
selected = wordvecselector.selectCandidates(subs, 'lexmturk.txt',
    proportion=0.75, stop_words_file='stop_words.txt', onlyInformative=True,
    keepTarget=True, onePerWord=True)
```

9.2 BiranSelector

Employs the strategy described in [2], in which a word co-occurrence model is used to determine which substitutions have meaning similar to that of a target complex word. It filters all substitutions which are estimated to be more complex than the target word, and also all those for which the distance between its co-occurrence vector and the target sentence's vector is higher than a threshold set by the user.

9.2.1 Parameters

To create a BiranSelector object, you must provide a word co-occurrence model. The model must be in plain text format, and each line must follow the format illustrated in Example 10.1, where $\langle w_i \rangle$ is a word, $\left\langle c_i^j \right\rangle$ a co-occurring word and $\left\langle f_i^j \right\rangle$ its frequency of appearance.

$$\langle w_i \rangle \ \left\langle c_i^0 \right\rangle : \left\langle f_i^0 \right\rangle \ \left\langle c_i^1 \right\rangle : \left\langle f_i^1 \right\rangle \cdots \left\langle c_i^{n-1} \right\rangle : \left\langle f_i^{n-1} \right\rangle \ \left\langle c_i^n \right\rangle : \left\langle f_i^n \right\rangle \tag{9.1}$$

Each component in the format above must be separated by a tabulation marker. To create a co-occurrence model, either create a script that does so, or follow the steps below:

1. Gather a corpus of text composed of one tokenized and truecased sentence per line.
2. Run the script resources/scripts/Produce_Co-occurrence_Model.py with the following command line:

   ```
   python Produce_Co-occurrence_Model.py <corpus> <window> <model_path>
   ```

   Where "<window>" is the number of tokens to the left and right of a word to be included as a co-occurring word.

To produce models faster, you can split your corpus in various small portions, run parallel processes to produce various small models, and then join them. For more information on how to use other class functions, please refer to LEXenstein's documentation.

9.2.2 Example

The code snippet below shows the BiranSelector class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.selectors import *
```

```
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

biranselector = BiranSelector('cooc_model.txt')
selected = biranselector.selectCandidates(subs, 'lexmturk.txt',
    common_distance=0.01, candidate_distance=0.9)
```

9.3 WSDSelector

Allows for the user to use many distinct classic WSD approaches in SS. It requires for the PyWSD [35] module to be installed, which includes the approaches presented by [21] and [36], as well as baselines such as random and first senses. The user can use any of the aforementioned approaches through the WSDSelector class by changing instantiation parameters.

9.3.1 Parameters

During instantiation, the WSDSelector class requires only for you to provide an id for the WSD method that you desire to use for Substitution Selection. For the options available, please refer to LEXenstein's documentation.

9.3.2 Example

The code snippet below shows the WSDSelector class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.spelling import *
```

```
nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

wsdselector = WSDSelector('lesk')
selected = wsdselector.selectCandidates(subs, 'lexmturk.txt')
```

9.4 AluisioSelector

Employs an SS strategy similar to the one introduced by [1]. It selects only those candidates which can assume the same POS tag as the target word. The selector initially parses a given sentence, and retrieves the POS tag of the target word. Using a POS tag conditional probability model, it then retrieves the set of possible tags of each candidate, and checks to see if the target's tag is in it.

9.4.1 Parameters

During instantiation, the AluisioSelector class requires for a POS tag conditional probability model, a binary POS tagging model, the path to a compiled version of the Stanford Tagger, and the path to the user's java installation.

The tagging model and tagger can both be downloaded from http://nlp.stanford.edu/software/tagger.shtml. The conditional probability model required must be created by the "createConditionalProbabilityModel" function from LEXenstein's Utilities module (lexenstein.utilities). For detailed instructions, please refer to the documentation of the "createConditionalProbabilityModel" function.

9.4.2 Example

The code snippet below shows the AluisioSelector class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.selectors import *
```

```
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

aluisioselector = AluisioSelector('cond_prob.bin', 'pos_model.tagger',
    'stanford-postagger.jar', '/usr/bin/java')
selected = aluisioselector.selectCandidates(subs, 'lexmturk.txt')
```

9.5 BelderSelector

Selects only those candidates which appear in the same word clusters in which a given target word is present. This strategy is inspired by the work of [10], in which synonyms are automatically extracted from a latent variable language model.

9.5.1 Parameters

During instantiation, the BelderSelector class requires for a file with clusters of words. The file must be in plain text format, and each line must follow the format illustrated in Example 9.2, which is the one adopted by the Brown Clusters implementation of [6]. In the Example 9.2, c_i is a class identifier, w_i a word, and f_i an optional frequency of occurrence of word w_i in the corpus over which the word classes were estimated.

$$\langle c_i \rangle \ \langle w_i \rangle \ \langle f_i \rangle \tag{9.2}$$

Each component in the format above must be separated by a tabulation marker. To create the file, one can use the software provided at https://github.com/percyliang/brown-cluster. Once the tool is installed, run the following command line:

```
./wcluster --text <corpus_of_sentences> --c <number_of_clusters>
```

The clusters file will be placed at **input-c50-p1.out/paths**.

9.5.2 Example

The code snippet below shows the BelderSelector class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

belderselector = BelderSelector('clusters.txt')
selected = belderselector.selectCandidates(subs, 'lexmturk.txt')
```

9.6 BoundarySelector

Employs a novel strategy, in which a Boundary Ranker is trained over a given set of features and then used to rank candidate substitutions according to how likely they are of being able to replace a target word without compromising the sentence's grammaticality or coherence. It retrieves a user-defined percentage of a set of substitutions.

9.6.1 Parameters

During instantiation, the BoundarySelector class requires for a BoundaryRanker object, which must be configured according to which features and resources the user intends to use to rank substitution candidates. The user can then use the "trainSelector" function to train the selector given a set of parameters, or the "trainSelectorWithCrossValidation" function to train it with cross-validation. Finally, the user can then retrieve a proportion of the candidate substitutions by using the "selectCandidates" function. For more information about the parameters of each function, please refer to LEXenstein's documentation.

9.6.2 Example

The code snippet below shows the BoundarySelector class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.features import *
from lexenstein.spelling import *
from lexenstein.rankers import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.txt', 0, 0, 'Complexity')
fe.addSenseCountFeature('Simplicity')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

br = BoundaryRanker(fe)
bs = BoundarySelector(br)
bs.trainSelectorWithCrossValidation('lexmturk.txt', 1, 5, 0.25)
selected = bs.selectCandidates(subs, 'lexmturk.txt', 'temp.txt', 0.25)
```

9.7 SVMBoundarySelector

Employs the same strategy used by the BoundaryRankerSelector, but instead of learning a linear model estimated over Stochastic Gradient Descent, it learns an either linear or non-linear model through Support Vector Machines. It retrieves a user-defined percentage of a set of substitutions.

9.7.1 Parameters

During instantiation, the SVMBoundarySelector class requires for a SVMBoundaryRanker object, which must be configured according to which features and resources the user intends to use to rank substitution candidates. The user can then use the "trainSelector" function to train the selector given a set of parameters, or the "trainSelectorWithCrossValidation" function to train it with cross-validation. Finally, the user can then retrieve a proportion of the candidate substitutions by using the "selectCandidates" function. For more information about the parameters of each function, please refer to LEXenstein's documentation.

9.7.2 Example

The code snippet below shows the SVMBoundarySelector class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.features import *
from lexenstein.spelling import *
from lexenstein.rankers import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')
fe.addSenseCountFeature('Simplicity')

br = SVMBoundaryRanker(fe)
sbs = SVMBoundarySelector(br)
sbs.trainSelectorWithCrossValidation('lexmturk.txt', 1, 5, 0.25)
selected = sbs.selectCandidates(subs, 'lexmturk.txt', 'temp.txt', 0.25)
```

9.8 SVMRankSelector

Employs a novel strategy, in which a SVM Ranker is trained over a given set of features and then used to rank candidate substitutions according to how likely they are of being able to replace a target word without compromising the sentence's grammaticality or coherence. It retrieves a user-defined percentage of a set of substitutions.

9.8.1 Parameters

During instantiation, the SVMRankSelector class requires for a SVMRanker object, which must be configured according to which features and resources the user intends to use to rank substitution candidates. The user can then use the "trainSelector" function to train the selector given a set of parameters, or the "trainSelectorWithCrossValidation" function to train it with cross-validation. Finally, the user can then retrieve a proportion of the candidate substitutions by using the "selectCandidates" function. For more information about the parameters of each function, please refer to LEXenstein's documentation.

9.8.2 Example

The code snippet below shows the SVMRankSelector class being used:

```
from lexenstein.morphadorner import MorphAdornerToolkit
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.features import *
from lexenstein.spelling import *
from lexenstein.rankers import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')
fe.addSenseCountFeature('Simplicity')
```

```
sr = SVMRanker(fe, './svm-rank/')
ss = SVMRankSelector(sr)
ss.trainSelectorWithCrossValidation('lexmturk.txt', 'f1.txt', 'm.txt', 5,
    0.25, './temp/', 0)
selected = ss.selectCandidates(subs, 'lexmturk.txt', 'f2.txt', 's1.txt',
    'temp.txt', 0.25)
```

9.9 VoidSelector

Does not perform any type of explicit substitution selection, and selects all possible substitutions generated for a given target word.

9.9.1 Parameters

During instantiation, the VoidSelector class requires no parameters.

9.9.2 Example

The code snippet below shows the VoidSelector class being used:

```
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.spelling import *

nc = NorvigCorrector('spelling_model.bin', format='bin')

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

voidselector = VoidSelector()
selected = voidselector.selectCandidates(subs, 'lexmturk.txt')
```

Chapter 10

The Substitution Ranking Module

Substitution Ranking (SR) is the task of ranking a set of selected substitutions for a target complex word with respect to its simplicity. Approaches vary from simple word length and frequency-based measures [2, 8, 9, 11] to more sophisticated linear combination as scoring functions [16] and Machine Learning approaches [15].

LEXenstein's SR module (lexenstein.rankers) provides access to 6 approaches, each represented by a Python class. The following Sections describe each one individually.

10.1 MetricRanker

Employs a simple ranking strategy based on the values of a single feature provided by the user. By configuring the input FeatureEstimator object, the user can calculate values of several features for the candidates in a given dataset, and easily rank the candidates according to each of these features.

10.1.1 Parameters

During instantiation, the MetricRanker requires only a configured FeatureEstimator object, which must contain at least one feature that can be used as a metric for ranking. Once created, the user can retrieve rankings by using MetricRanker's "getRankings" function, which requires a VICTOR corpus and the index of a feature to be used as a ranking metric.

10.1.2 Example

The code snippet below shows the MetricRanker class being used:

```
from lexenstein.features import *
```

```
from lexenstein.rankers import *

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')
fe.addLengthFeature('Complexity')
fe.addSenseCountFeature('Simplicity')

mr = MetricRanker(fe)
frequency_ranks = mr.getRankings('lexmturk.txt', 0)
length_ranks = mr.getRankings('lexmturk.txt', 1)
sense_ranks = mr.getRankings('lexmturk.txt', 2)
```

10.2 SVMRanker

Use Support Vector Machines in a setup that minimizes a loss function with respect to a ranking model. In LS, this strategy is the one employed in the experiments of [15], yielding promising results.

10.2.1 Parameters

During instantiation, the SVMRanker requires for a FeatureEstimator object and a path to the root installation folder of SVM-Rank [17]. The user can then use the "getFeaturesFile", "getTrainingModel", "getScoresFile" and "getRankings" to train SVM ranking models and rank candidate substitutions. For more information on these functions' parameters, please refer to LEXenstein's documentation and the example in the following Section.

10.2.2 Example

The code snippet below shows the SVMRanker class being used:

```
from lexenstein.features import *
from lexenstein.rankers import *

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')
fe.addLengthFeature('Complexity')
fe.addSenseCountFeature('Simplicity')
```

```
svmr = SVMRanker(fe, '/svm-rank/')
svmr.getFeaturesFile('lexmturk.txt', 'features.txt')
svmr.getTrainingModel('features.txt', 0.1, 0.1, 0, 'model.txt')
svmr.getScoresFile('features.txt', 'model.txt', 'scores.txt')
rankings = svmr.getRankings('features.txt', 'scores.txt')
```

10.3 BoundaryRanker

Employs a novel strategy, in which ranking is framed as a binary classification task. During training, this approach assigns the label 1 to all candidates of rank $1 \geq r \geq p$, where p is a range set by the user, and 0 to the remaining candidates. It then trains a stochastic descent linear classifier based on the features specified in the FeatureEstimator object. During testing, candidate substitutions are ranked based on how far from 0 they are.

10.3.1 Parameters

During instantiation, the BoundaryRanker requires only for a FeatureEstimator object. The user can then use the "trainRanker" function to train a ranking model, and the "getRankings" to rank the candidates of a VICTOR corpus. For more information on the training parameters supported by the "trainRanker" function, please refer to LEXenstein's documentation.

10.3.2 Example

The code snippet below shows the BoundaryRanker class being used:

```
from lexenstein.features import *
from lexenstein.rankers import *

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')
fe.addLengthFeature('Complexity')
fe.addSenseCountFeature('Simplicity')

br = BoundaryRanker(fe)
br.trainRanker('lexmturk.txt', 1, 'modified_huber', 'l1', 0.1, 0.1, 0.001)
rankings = br.getRankings('lexmturk.txt')
```

10.4 SVMBoundaryRanker

Employs the same strategy used by the BoundaryRanker class, but instead of learning a linear ranking model through Stochastic Gradient Descent, it learns a linear or non-linear model by using Support Vector Machines.

10.4.1 Parameters

During instantiation, the SVMBoundaryRanker requires only for a FeatureEstimator object. The user can then use the "trainRanker" function to train a ranking model, and the "getRankings" to rank the candidates of a VICTOR corpus. For more information on the training parameters supported by the "trainRanker" function, please refer to LEXenstein's documentation.

10.4.2 Example

The code snippet below shows the SVMBoundaryRanker class being used:

```
from lexenstein.features import *
from lexenstein.rankers import *

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')
fe.addLengthFeature('Complexity')
fe.addSenseCountFeature('Simplicity')

sbr = SVMBoundaryRanker(fe)
sbr.trainRanker('lexmturk.txt', 1, 10, 'poly', 2, 0.1, 1)
rankings = sbr.getRankings('lexmturk.txt')
```

10.5 BiranRanker

Employs the strategy of [2], which models complexity as a function of a word's length and frequency in corpora of complex and simple content. As input, it requires for a language model trained over complex data, and a language model trained over simple data.

10.5.1 Parameters

The language models required by the BiranRanker must be binary, and must be produced by KenLM with the following command lines:

```
lmplz -o [order] <[corpus_of_text] >[language_model_name]
```

```
build_binary [language_model_name] [binary_language_model_name]
```

Complex and simple data can be downloaded fro David Kauchak's page[1], or extracted from other sources.

10.5.2 Example

The code snippet below shows the BiranRanker class being used:

```
from lexenstein.rankers import *

br = BiranRanker('lm_complex.bin', 'lm_simple.bin')
rankings = br.getRankings('lexmturk.txt')
```

10.6 YamamotoRanker

Employs the strategy of [19], which ranks words according to a weighted function that considers their frequency in a corpus of simple content, co-occurrence frequency with a target complex word, sense distance, point-wise mutual information and trigram frequencies. As input, it requires for a language model trained over a corpus of simple data and a co-occurrence model.

10.6.1 Parameters

The language model required by the YamamotoRanker must be binary, and must be produced by KenLM with the following command lines:

```
lmplz -o [order] <[corpus_of_text] >[language_model_name]
```

```
build_binary [language_model_name] [binary_language_model_name]
```

[1] http://www.cs.pomona.edu/ dkauchak/simplification/

Complex and simple data can be downloaded fro David Kauchak's page[2], or extracted from other sources. The co-occurrence model must be in plain text format, and each line must follow the format illustrated in Example 10.1, where $\langle w_i \rangle$ is a word, $\left\langle c_i^j \right\rangle$ a co-occurring word and $\left\langle f_i^j \right\rangle$ its frequency of appearance.

$$\langle w_i \rangle \ \langle c_i^0 \rangle : \langle f_i^0 \rangle \ \langle c_i^1 \rangle : \langle f_i^1 \rangle \cdots \langle c_i^{n-1} \rangle : \langle f_i^{n-1} \rangle \ \langle c_i^n \rangle : \langle f_i^n \rangle \tag{10.1}$$

Each component in the format above must be separated by a tabulation marker. To create a co-occurrence model, either create a script that does so, or follow the steps below:

1. Gather a corpus of text composed of one tokenized and truecased sentence per line.

2. Run the script resources/scripts/Produce_Co-occurrence_Model.py with the following command line:

```
python Produce_Co-occurrence_Model.py <corpus> <window> <model_path>
```

   Where "<window>" is the number of tokens to the left and right of a word to be included as a co-occurring word.

To produce models faster, you can split your corpus in various small portions, run parallel processes to produce various small models, and then join them. For more information on the parameters of each function of the YamamotoRanker class, please refer to the LEXenstein documentation.

10.6.2 Example

The code snippet below shows the YamamotoRanker class being used:

```
from lexenstein.rankers import *

yr = YamamotoRanker('lm_simple.bin', 'cooc_model.txt')
rankings = yr.getRankings('lexmturk.txt')
```

[2]http://www.cs.pomona.edu/ dkauchak/simplification/

10.7 BottRanker

Employs the strategy of [5], which ranks words according to a weighted function that considers their length and frequency in a corpus of simple content. As input, it requires for a language model trained over a corpus of simple data.

10.7.1 Parameters

The language model required by the BottRanker must be binary, and must be produced by KenLM with the following command lines:

```
lmplz -o [order] <[corpus_of_text] >[language_model_name]
```

```
build_binary [language_model_name] [binary_language_model_name]
```

Complex and simple data can be downloaded fro David Kauchak's page[3], or extracted from other sources.

10.7.2 Example

The code snippet below shows the BottRanker class being used:

```
from lexenstein.rankers import *

br = BottRanker('lm_simple.bin')
rankings = br.getRankings('lexmturk.txt')
```

10.8 GlavasRanker

Employs the strategy of [13], which ranks words according to their average ranking, as determined by a given set of features. As input, it requires for a configured FeatureEstimator object.

10.8.1 Parameters

The FeatureEstimator object required must contain each and every feature that the user wishes to rank candidates with.

[3] http://www.cs.pomona.edu/ dkauchak/simplification/

10.8.2 Example

The code snippet below shows the GlavasRanker class being used:

```
from lexenstein.features import *
from lexenstein.rankers import *

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')
fe.addLengthFeature('Complexity')
fe.addSenseCountFeature('Simplicity')

gr = GlavasRanker(fe)
rankings = gr.getRankings('lexmturk.txt')
```

Chapter 11

The Evaluation Module

Since one of the goals of LEXenstein is to facilitate the benchmarking LS approaches, it is crucial that it provides evaluation methods. LEXenstein's evaluation module (lexenstein.evaluators) includes functions for the evaluation of all sub-tasks, both individually and in combination. It contains 5 classes, each designed for one form of evaluation. We discuss them in more detail in the following Sections.

11.1 IdentifierEvaluator

Provides evaluation metrics for CWI methods. It requires a gold-standard in the CWICTOR format and a set of binary word complexity labels. The labels must have value 1 for complex words, and 0 otherwise. It returns the Accuracy, Precision, Recall, F-score and G-score, which is the harmonic mean between Accuracy and Recall.

The code snippet below shows the IdentifierEvaluator class being used:

```
from lexenstein.identifiers import *
from lexenstein.evaluators import *

li = LexiconIdentifier('lexicon.txt', 'simple')
labels = li.identifyComplexWords('test_cwictor_corpus.txt')

ie = IdentifierEvaluator()
accuracy, precision, recall, fmean, gmean =
    ie.evaluateIdentifier('test_cwictor_corpus.txt', labels)
```

11.2 GeneratorEvaluator

Provides evaluation metrics for SG methods. It requires a gold-standard in the VICTOR format and a set of generated substitutions. It returns the Potential, Precision, Recall and F-measure, where Potential is the proportion of instances in which at least one of the substitutions generated is present in the gold-standard, Precision the proportion of generated instances which are present in the gold-standard, Recall the proportion of gold-standard candidates that were generated, and F-measure the harmonic mean between Precision and Recall.

The code snippet below shows the GeneratorEvaluator class being used:

```
from lexenstein.spelling import *
from lexenstein.generators import *
from lexenstein.evaluators import *
from lexenstein.morphadorner import *

m = MorphAdornerToolkit('./morph/')

nc = NorvigCorrector('spelling_model.bin', format='bin')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

ge = GeneratorEvaluator()
potential, precision, recall, fmeasure =
    ge.evaluateGenerator('lexmturk.txt', subs)
```

11.3 SelectorEvaluator

Provides evaluation metrics for SS methods. It requires a gold-standard in the VICTOR format and a set of selected substitutions. It returns the Potential, Precision and F-measure of the SS approach, where Potential is the proportion of instances in which at least one of the substitutions selected is present in the gold-standard, Precision the proportion of selected candidates which are present in the gold-standard, Recall the proportion of gold-standard candidates that were selected, and F-measure the harmonic mean between Precision and Recall.

The code snippet below shows the SelectorEvaluator class being used:

```
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.evaluators import *
from lexenstein.morphadorner import *

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

biranselector = BiranSelector('cooc_model.txt')
selected = biranselector.selectCandidates(subs, 'lexmturk.txt', 0.01, 0.75)

se = SelectorEvaluator()
potential, precision, recall, fmeasure = se.evaluateSelector('lexmturk.txt',
    selected)
```

11.4 RankerEvaluator

Provides evaluation metrics for SR methods. It requires a gold-standard in the VICTOR format and a set of ranked substitutions. It returns the TRank-at-1 : 3 and Recall-at-1 : 3 metrics [34], where Trank-at-i is the proportion of instances in which a candidate of gold-rank $r \leq i$ was ranked first, and Recall-at-i the proportion of candidates of gold-rank $r \leq i$ that are ranked in positions $p \leq i$.

The code snippet below shows the RankerEvaluator class being used:

```
from lexenstein.rankers import *
from lexenstein.features import *

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')

mr = MetricRanker(fe)
rankings = mr.getRankings('lexmturk.txt', 0)
```

```
re = RankerEvaluator()
t1, t2, t3, r1, r2, r3 = re.evaluateRanker('lexmturk.txt', rankings)
```

11.5 PipelineEvaluator

Provides evaluation metrics for the entire LS pipeline. It requires as input a gold-standard in VICTOR format and a set of ranked substitutions which have been generated and selected by a given set of approaches. It returns the approaches' Precision, Accuracy and Change Proportion, where Precision is the proportion of instances in which the highest ranking substitution is not the target complex word itself and is in the gold-standard, Accuracy is the proportion of instances in which the highest ranking substitution is in the gold-standard, and Change Proportion is the proportion of instances in which the highest ranking substitution is not the target complex word itself.

The code snippet below shows the PipelineEvaluator class being used:

```
from lexenstein.generators import *
from lexenstein.selectors import *
from lexenstein.evaluators import *
from lexenstein.rankers import *
from lexenstein.features import *
from lexenstein.morphadorner import *

m = MorphAdornerToolkit('./morph/')

wg = WordnetGenerator(m, nc, 'pos_model.tagger', 'stanford-postagger.jar',
    '/usr/bin/java')
subs = wg.getSubstitutions('lexmturk.txt')

bs = BiranSelector('cooc_model.txt')
selected = bs.selectCandidates(subs, 'lexmturk.txt', 0.01, 0.75)
bs.toVictorFormat('lexmturk.txt', selected, 'victor.txt',
    addTargetAsCandidate=True)

fe = FeatureEstimator()
fe.addCollocationalFeature('lm.bin', 0, 0, 'Complexity')

mr = MetricRanker(fe)
```

```
rankings = mr.getRankings('victor.txt', 0)

pe = PipelineEvaluator()
precision, accuracy, changed = pe.evaluatePipeline('lexmturk.txt', rankings)
```

References

[1] Aluisio, S. and Gasperin, C. (2010). *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, chapter Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts, pages 46–53. Association for Computational Linguistics.

[2] Biran, O., Brody, S., and Elhadad, N. (2011). Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th ACL*, pages 496–501.

[3] Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly.

[4] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*.

[5] Bott, S., Rello, L., Drndarevic, B., and Saggion, H. (2012). Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of CoLing*, pages 357–374.

[6] Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

[7] Burns, P. R. (2013). MorphAdorner v2: A Java Library for the Morphological Adornment of English Language Texts.

[8] Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

[9] Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL*, pages 269–270.

[10] De Belder, J. and Moens, M.-F. (2010). Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

[11] Devlin, S. and Tait, J. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

[12] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

[13] Glavaš, G. and Štajner, S. (2015). Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd ACL*.

[14] Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

[15] Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd ACL*, pages 458–463.

[16] Jauhar, S. and Specia, L. (2012). UOW-SHEF: SimpLex–lexical simplicity ranking based on contextual and psycholinguistic features. pages 477–481.

[17] Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM*, pages 133–142.

[18] Jones, E., Oliphant, T., Peterson, P., et al. (2015). SciPy: Open source scientific tools for Python.

[19] Kajiwara, T., Matsumoto, H., and Yamamoto, K. (2013). Selecting Proper Lexical Paraphrase for Children. *Proceedings of the 25th Rocling*, pages 59–73.

[20] Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st ACL*, pages 423–430.

[21] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Conference on Systems Documentation*, pages 24–26.

[22] Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19:313–330.

[23] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[24] Nunes, B. P., Kawase, R., Siehndel, P., Casanova, M. a., and Dietze, S. (2013). As Simple as It Gets - A Sentence Simplifier for Different Learning Levels and Contexts. pages 128–132.

[25] Paetzold, G. H. (2015). Morph adorner toolkit: Morph adorner made simple. http://ghpaetzold.github.io/MorphAdornerToolkit/.

[26] Paetzold, G. H. and Specia, L. (2013). Text simplification as tree transduction. In *Proceedings of the 9th STIL*.

[27] Paetzold, G. H. and Specia, L. (2016). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th SemEval*.

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[29] Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA.

[30] Sedding, J. and Kazakov, D. (2004). Wordnet-based text document clustering. In *Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data*, ROMAND '04, pages 104–113, Stroudsburg, PA, USA. Association for Computational Linguistics.

[31] Shardlow, M. (2013a). A comparison of techniques to automatically identify complex words. pages 103–109.

[32] Shardlow, M. (2013b). *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, chapter The CW Corpus: A New Resource for Evaluating the Identification of Complex Words, pages 69–77. Association for Computational Linguistics.

[33] Shelley, M. (2007). *Frankenstein*. Pearson Education.

[34] Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 1st SemEval*, pages 347–355.

[35] Tan, L. (2014). Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]. https://github.com/alvations/pywsd.

[36] Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of ACL*.