ColLex.EN: Automatically Generating and Evaluating a Full-form Lexicon for English

Tim vor der Brück, Alexander Mehler and Zahurul Islam

Text-Technology Lab
Computer Science
Goethe- University Frankfurt

Abstract

The paper describes a procedure for the automatic generation of a large full-form lexicon of English. We put emphasis on two statistical methods to lexicon extension and adjustment: in terms of a letter-based HMM and in terms of a detector of spelling variants and misspellings. The resulting resource, ColLex.EN, is evaluated with respect to two tasks: text categorization and lexical coverage by example of the SUSANNE corpus and the Open ANC.

Keywords: Lexical ressource, lexicon generation, text classification

1. Introduction

Currently, a large number of different lexica is available for English. However, substantial and freely available full-form lexica with a high number of named entities are rather rare even in the case of this lingua franca. Existing lexica are often limited in several respects as explained in Section 2. What is missing so far is a freely available substantial machine-readable lexical resource of English that contains a high number of word forms and a large collection of named entities. In this paper, we describe a procedure to generate such a resource by example of English. This lexicon, henceforth called *ColLex.EN* (for *Collecting Lexica for English*), will be made freely available to the public[1]. In this paper, we describe how ColLex.EN was collected from existing lexical resources and specify the statistical procedures that we developed to extend and adjust it. No manual modifications were done on the generated word forms and lemmas. Our fully automatic procedure has the advantage that whenever new versions of the source lexica are available, a new version of ColLex.EN can be automatically generated with low effort.

2. Related Work

Since lexical resources are one of the most important components of many NLP applications, the effort of merging existing resources to produce a richer one is always an issue. Crouch and King (2005) describe merging lexical resources in two steps. The first step consists of extracting lexical entries from different resources and mapping them into a common format. Then, all the entries are compared automatically and combined to form the merged resource.

Bel et al. (2011) follow the lexical merging approach described in Crouch and King (2005). The process consist of writing entry extraction rules manually and then using a graph-based unification tool (Bird, 2006) to unify handwritten sub-categorization information.

Gurevych et al. (2012a) present UBY, a very good resource for German and English, which results from merging nine existing resources of these two languages (Gurevych et al., 2012b). This is done by means of an extended version of LMF to accommodate entries from heterogeneous resources as, for example, WordNet, VerbNet (Kipper et al., 2008) and GermaNet (Kunze and Lemnitzer, 2002). The English version of UBY contains a large number of named entities, but only few word forms. Further, a lot of named entities are marked as common nouns and links from past tense forms to present tense forms are missing.

Bhatt et al. (2013) present IndoNet, a multilingual lexical resource, generated by merging various existing resources such as the *universal word dictionary* (Uchida et al., 2000), SUMO (Niles and Pease, 2001), English and Hindi WordNet. IndoNet uses a common concept hierarchy to link these resources.

Another multilingual resource, which is also available for English, is BabelNet (Navigli and Ponzetto, 2012; Navigli, 2013). It is one of the largest existing multilingual lexical resources and has been generated, amongst others, by automatically integrating the English *WordNet* and *Wikipedia*. BabelNet is a graph-based resource where each vertex denotes a concept and each edge represents a semantic relation among the connected concepts. Senses from *WordNet* and pages from *Wikipedia* are automatically mapped to generate the resource. The latest version of BabelNet contains 44 482 718 lemmas from 51 languages. This includes a vast number of named entities but misses to map their different word forms. Further, many proper nouns covered by BabelNet actually belong to another language (e.g., Chinese) and not to the indicated one (i.e., English).

In contrast to BabelNet, WordNet (Fellbaum, 1998) contains plenty of regular and irregular word forms, but only a small number of named entities. Obviously, the latter deficit does not hold for Wikipedia that is mainly intended to be used by human readers. Wikipedia is not a formally structured resource so that a lot of effort has to be spent to parse it.

Most of the resources mentioned so far focus more on quantity. As a result, word forms are often missed in the lexicon. Further, spelling variants or misspellings (as often found in texts) are not resolved and mapped onto their correct lemma. On the other hand, high quality resources like

[1] See `http://english.lexicon.hucompute.org`

WordNet do not contain large lists of proper nouns. In this paper, we describe a procedure to generate ColLex.EN as a collected lexical resource of English that aims at overcoming these deficits.
The paper is organized as follows: Section (3.) describes the different resources that we explored to build ColLex.EN and the basic procedure to generate it. Section (3.1.) describes our procedure for the detection of variants and misspellings. Section (4.) presents statistics on ColLex.EN. Finally, Section (5.) presents a comparative evaluation of our resource.

3. Generating ColLex.EN

The basic procedure of generating ColLex.EN consists of three steps: (1) Merging the input lexica, (2) implementing a HMM for grapheme-based PoS-tagging of unknown words, and (3) unifying spelling variants by means of a Levenshtein metric that operates on a skip-tolerant trie. The latter two steps are performed to extend ColLex.EN beyond the input resources. For this task we use the complete English Wikipedia and try to process any input token that is unknown from the point of view of the input resources. Since this procedure detects spelling variants and misspellings, it allows for a controlled extension of the set of word forms in the lexicon.
The following resources were explored during the merging step (Step 1):

1. The ClearNLP[2] project which provides a lexicon together with word formation rules that we used to generate a full-form lexicon.
2. The English Wiktionary, for which we implemented an novel extractor.
3. The English Wikipedia, for which we implemented a parser extractor.
4. The Hunspell lexicon[3], which, amongst others, is exploited by LibreOffice. Since the coverage of Hunspell is small we applied the ClearNLP rules to generate additional word forms.
5. The UBY[4] lexicon for English (Gurevych et al., 2012b).
6. UBY+: a variant of UBY that we generated by applying the word formation rules of ClearNLP to UBY.
7. WordNet (Fellbaum, 1998).
8. A list of proper nouns of the German National Library[5] that we parsed for prenames and last names.
9. The lexicon of the Preprocessor2010 (Waltinger, 2010) which, amongst others, is based on the CELEX.

In any of these cases, WordNet is the reference in the merging process. For each input word that has to be merged, we save information about all source lexica considered here in which it was found. We also automatically corrected entries of the input lexica by exploiting knowledge about their shortcomings (e.g., incorrect assignments of common nouns and proper nouns, wrong capitalization). The procedure for differentiating common nouns and proper nouns is based on WordNet and corpus statistics. First, we verify whether the input word is contained in WordNet. If this is the case, we check whether the word has a sense that is marked as an *instance*. If this is the case, the word is tagged as a proper noun, otherwise as a common noun. If the word does not belong to WordNet, we exploit the fact that proper nouns are capitalized in English. Thus, if the word is capitalized more often than not within our reference corpus (i.e., Wikipedia), it is marked as an proper noun.

Further, in order to prevent that malformed or wrongly tagged input of the reference corpora is added to ColLex.EN (in the lexicon of the Preprocessor2010, for example, *yeastiness* is wrongly tagged as an adverb, though by its ending it can be detected as a noun), we apply several linguistic and statistical verification methods when processing lexemes of open word classes. Note that these methods are only applied to units that are probably not named entities. The reason is that we do not have a consistent set of morphological rules for detecting named entities of any sort. For named entities, we only check whether they consist of Latin characters. (Thus, words in Chinese letters, for example, are not included into ColLex.EN.)

Our first and simplest validation method is to look up the input string in WordNet. If this is successful, we insert the word in our lexicon. If this is not the case we lookup constituents of the word in WordNet. If a word consists of several parts that are either separated by a space or a hyphen, only the last part of the word is required to be contained in WordNet. If this is not the case, we further check whether the word contains a suffix that indicates a certain part of speech (e.g., *-ion* in the case of nouns). If this test also fails (and only then), a statistical validation method is applied. This includes the application of a *Hidden Markov Model* (HMM) based on letter tetragrams to determine the most likely PoS of the input word. Two special characters are introduced to indicate the beginning and end of a word. Four independent HMMs are trained for each of the considered parts of speech (*adjective*, *adverb*, *noun* and *verb*). Generally speaking, the probability for the observed state-sequence is given by:

$$\begin{aligned} & P(X_1 = x_1, \ldots, X_n = x_n) \\ =& P(X_1 = x_1) P(X_2 = x_2 | X_1 = x_1) \cdot \\ & P(X_3 | X_1 = x_1, X_2 = x_2) \cdot \ldots \cdot \\ & P(X_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \end{aligned} \quad (1)$$

where x_i is the ith character of the considered word. Following the Markov assumption that the occurrence of a character in a word only depends on a fixed number of pre-

[2] code.google.com/p/clearnlp
[3] hunspell.sourceforge.net
[4] www.ukp.tu-darmstadt.de/data/lexical-resources/uby
[5] www.dnb.de/EN/Home/home_node.html

ceding characters (here 4) and taking logs we get:

$$\begin{aligned} &\ln P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \sum_{i=1,\ldots,n} \ln(P(X_i = x_i | X_{P(i)} = x_{p(i)}) \end{aligned} \quad (2)$$

where X_i denotes the *i*th character of the input word. $X_{p(i)}$ and $x_{p(i)}$ is a sequence of predecessor variables of X_i and its associated values with a length of at most 4. For training we used all lemmas contained in WordNet. We estimated the logarithmized probabilities that the word is created by each of the models and chose the PoS for which the associated model produced the highest probability. To account for missing data, we used a smoothing method regarding unigrams, bigrams, trigrams and tetragrams. Further, we derived a normalized probability estimate (to provide independence of word length). It is given by the geometric mean of the multiplied probabilities:

$$\begin{aligned} &\ln((\prod_{i=1}^{n}(P(X_i = x_i | X_{p(i)} = x_{p(i)})))^{1/n}) \\ &= (1/n) \sum_{i=1}^{n} \ln(P(X_i = x_i | X_{p(i)} = x_{p(i)})) \end{aligned} \quad (3)$$

This normalized estimate allows for comparing words of different length. If the logarithmized normalized probability of the most likely PoS-hypothesis is smaller than the 2%-quantile, the word is rejected as not belonging to any of the considered categories. In this case, we did not include it in ColLex.EN unless a WordNet-based or morphological validation was applicable (for example, by testing whether the word ends by a typical suffix of the PoS under consideration). A word is also not included in ColLex.EN in cases where the PoS determined by the HMM does not coincide with the PoS annotated in the input resource. Extending ColLex.EN in future work will operate on these cases by considering words whose PoS is probably better determined by our HMM. Currently, we aim at achieving a higher precision at cost of a higher coverage.

3.1. Automatic Detection of Variants

An integral part of building ColLex.EN is to extract and process spelling variants found in Wikipedia and to add them to ColLex.EN whenever possible. To this end, all lemmas of ColLex.EN were stored in a letter trie. Then, we looked up all unknown word forms in Wikipedia in the trie. The lookup is error-tolerant (Eisele and vor der Brück, 2004) up to a certain number of errors using a Levenshtein metric. However, for the first characters of the input string we required a perfect match in the trie. One reason is that spelling variants or misspellings are less likely at the beginning of a word. At each character position in the input we look for the child node of the current trie node that is associated to this character, beginning with the first character of the input and the root node of the trie. If the character is found, we move to the associated child node. Note that we additionally allow for skipping, inserting, replacing, and exchanging characters in the input string. A skip relates to the case that we move forward by one character in the input string but stay on the current node in the trie. If we insert one character from the trie to the input string, we do not move forward in the input but step forward in the trie. In addition to the original Levenshtein metric (Levenshtein, 1966) we allow for permuting two neighboring characters in the input. Table 1 exemplifies variants (left) and their lemmas (right) found by this procedure.

Variant	Lemma
Aaalborg	Aalborg
Aaarhus	Aarhus
Abchazia	Abkhazia

Table 1: Sample variants and their lemmas.

PoS	Number of word forms
verb	150 172
noun	1 129 183
proper noun	10 306 741
adjective	285 528
adverb	15 382
other word form	20 669
total	11 907 675

Table 2: Word forms contained in ColLex.EN.

4. Statistics about ColLex.EN

Table 2 shows statistics of ColLex.EN as generated by the procedure described in Section 3. Each entry consists of the word form, the lemma, its part of speech and the names of all resources (methods) that the entry was extracted from (generated by). It additionally contains several lists of named entities of different types (e.g., first names, second names, locations, planets, etc.). ColLex.EN will be made available according to LMF[6].[7]

5. Evaluation

We evaluated ColLex.EN in comparison to its input resources. Our hypothesis was that ColLex.EN performs at least as good as these input resources.

The evaluation was done by means of two tasks. We started with performing a DDC-related multi-label text categorization (Mehler and Waltinger, 2009; Waltinger et al., 2011). DDC is the abbreviation of *Dewey Document Categorization* and denotes a hierarchical categorization scheme mainly employed by digital libraries. DDC contains ten top-level categories that we used as target classes. Text categorization (Joachims, 2002) is the task of automatically assigning texts onto a set of predefined categories, for example, of *genre* or *topic*. Our categorization experiment is based on a supervised machine learning approach using support vector machines that are trained by means of a subset of documents for which correct topics are known independently (Lösch et al., 2011). Our classifier is publicly available via its website: `ddc.hucompute.org`.

[6] `www.lexicalmarkupframework.org`

[7] See `http://english.lexicon.hucompute.org`

Lexicon	F-measure	Precision	Recall
ColLex.EN	0.735	0.725	**0.770**
WordNet	**0.737**	**0.731**	0.765
UBY	0.715	0.702	0.759
Wiktionary	0.724	0.723	0.727
BabelNet	0.661	0,693	0,680

Table 3: Results of the text categorization experiment.

For the first (categorization) task, we determined overall precision, recall and F-measures by doing a macro-averaging over all categories (see Table 3). The test set consists of three samples of 500 abstracts of the English OAI (*Open Archive Initiative*) corpus. For the second task we evaluated ColLex.EN in comparison to its underlying resources (see Section 3.) regarding the coverage of words found in two reference corpora. In the latter task, we used the freely available SUSANNE corpus (Sampson, 1995) and the Open ANC (Ide and Suderman, 2004) to answer the question how many of the lemmas found in these corpora are known by the lexica and whether these resources allow for tagging the PoS of these lemmas correctly. The SUSANNE corpus consists of approximately 100 000 words and is a subset of the Brown Corpus of American English. The Open ANC consists of 14 million words of spoken data and written texts. Each token of the SUSANNE and the Open ANC is tagged by its lemma and PoS. While the Open ANC is automatically tagged by a part of speech tagger, the annotations within the SUSANNE corpus where done manually. For our evaluation, we used the entire SUSANNE corpus and the slate subcorpus of Open ANC with around four million words. We determined for each token, whether the associated lemma and PoS, as annotated in the input corpus, is covered by the corresponding lexicon to be evaluated and calculated the accuracy (see Table 4 and 5). A token was considered to be correctly recognized, if the annotated lemma was found in the lexicon for the associated word form (part of speech tags were evaluated analogously).

The evaluation on the two test corpora shows that ColLex.EN reaches the highest accuracy on both input corpora, followed in both cases by UBY+. Though the Preprocessor2010 reaches an accuracy of lemmatization even higher than the one of UBY+, its PoS accuracy is much worse. Note also that ColLex.EN outperforms all other lexica in terms of the PoS-related coverage test. However, the use of ColLex.EN resulted only in the second best F-score in the text categorization task. Here, WordNet produces the highest score directly followed by ColLex.EN, which produces the highest recall among all resources considered here.

6. Error Analysis and Discussion

The manually annotated SUSANNE corpus is of high quality. This does not hold for the automatically tagged Open ANC with respect to PoS tagging. Note also that the Open ANC contains several mistakes. A lemma with suffix 'um', for example, is often created for word forms that ended by '*a*' (e.g., *Paula* is mapped to the lemma *paulum*). In addition, lemmas are written in lowercases, even named entities.

Lexicon	Lemma accuracy	PoS accuracy
ColLex.EN	**0.976**	**0.678**
Preprocessor	0.938	0.405
UBY+	0.907	0.645
Wiktionary	0.860	0.415
UBY	0.813	0.644
BabelNet	0.743	0.321
WordNet	0.683	0.405
ClearNLP	0.678	0.403
Hunspell	0.590	—
Wikipedia	0.565	—

Table 4: Lexicon coverage test by example of the SUSANNE corpus. Note that in all coverage tests, we applied the ClearNLP rules whenever the respective resource only provided lemma information (as in the case of BabelNet). Order according to the accuracy of lemmatization.

Lexicon	Lemma accuracy	PoS accuracy
ColLex.EN	**0.972**	**0.871**
Preprocessor	0.925	0.507
UBY+	0.916	0.776
BabelNet	0.896	0.407
Wiktionary	0.857	0.472
UBY	0.831	0.773
ClearNLP	0.678	0.450
WordNet	0.660	0.464
Hunspell	0.600	—
Wikipedia	0.578	—

Table 5: Lexicon coverage test by example of the Open ANC.

It was not possible to correct all these errors automatically so that the coverage test needed to work on this somehow erroneous resource. In the case of the Open ANC, this may explain the rates documented in Table 5. However, since all lexical resources considered here work on an equal footing what regards these errors, their order relation as documented in Table 4 and 5 is possibly not affected.

Regarding our experiment on text categorization, we see that a larger resource does not necessarily produce a better result. *Why?* ColLex.EN comes with a huge list of proper nouns that contains, for example, names like *Seventeen* or *All*. The use of this list may result in wrong output of the preprocessor in the sense that insignificant lexical features are wrongly attributed to be relevant. If this is a correct error analysis than we may resume: *In case of classification tasks, the larger the underlying lexical resource, the more important the process of feature selection.*

7. Conclusion

We presented a procedure for automatically generating a large lexical resource of English. This has been done by combining several existing lexica. In order to prevent malformed input, we applied several statistical verification methods. As a result of the verification process, lexical entries were filtered out or modified automatically and stored in the final lexicon ColLex.EN. We evaluated our newly generated lexicon ColLex.EN in comparison to several lex-

ical resources. This has been done by example of two tasks: *text categorization* and *lexical coverage* regarding two tagged corpora. The evaluation shows that ColLex.EN outperformed the other lexica in the coverage test and also achieved an F-score in the categorization task that is nearly as good as the best performing resource in this second test, that is, WordNet.

8. References

Núria Bel, Muntsa Padró, and Silvia Necsulescu. 2011. A method towards the fully automatic merging of lexical resources. In *Proceedings of the ACL Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 8–15.

Brijesh Bhatt, Lahari Poddar, and Pushpak Bhattacharyya. 2013. IndoNet: A multilingual lexical knowledge network for indian languages. In *51st Annual Meeting of the Association for Computational Linguistics*.

Steven Bird. 2006. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72.

Dick Crouch and Tracy Holloway King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Andreas Eisele and Tim vor der Brück. 2004. Error-tolerant finite-state lookup for trademark search. In *27th Annual German Conference on AI, KI 2004*, number 3238 in LNCS, pages 112–126.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database. Cambridge*. MIT Press.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012a. UBY – a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*.

Iryna Gurevych, Michael Matuschek, Tri Duc Nghiem, Judith Eckle-Kohler, Silvana Hartmann, and Christian M. Meyer. 2012b. Navigating sense-aligned lexical-semantic resources: The web interface to UBY. In *Proceedings of the 11th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, pages 194–198.

N. Ide and K. Suderman. 2004. The American National Corpus first release. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, pages 1681–1684, Lisbon, Portugal.

Thorsten Joachims. 2002. *Learning to classify text using support vector machines*. Kluwer, Boston.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

Claudia Kunze and Lothat Lemnitzer. 2002. GermaNet – representation, visualization, application. In M. Rodriguez González and C. Paz Suárez Araujo, editors, *Proceedings of LREC 2002*, pages 1485–1491, Paris.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Mathias Lösch, Ulli Waltinger, Wolfram Horstmann, and Alexander Mehler. 2011. Building a DDC-annotated corpus from OAI metadata. *Journal of Digital Information*, 12(2).

Alexander Mehler and Ulli Waltinger. 2009. Enhancing document modeling by means of open topic models: Crossing the frontier of classification schemes in digital libraries by example of the DDC. *Library Hi Tech*, 27(4):520–539.

R. Navigli and S. Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2013. A quick tour of BabelNet 1.1. In *Computational Linguistics and Intelligent Text Processing*, pages 25–37. Springer.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM.

G.R. Sampson. 1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford University Press.

Hiroshi Uchida, Meiying Zhu, and Tarcisio Della Senta. 2000. UNL: A gift for a millennium. Technical report, The United Nations University.

Ulli Waltinger, Alexander Mehler, Mathias Lösch, and Wolfram Horstmann. 2011. Hierarchical classification of OAI metadata using the DDC taxonomy. In *Advanced Language Technologies for Digital Libraries*, number 6699 in LNCS, pages 29–40. Springer, Heidelberg.

Ulli Waltinger. 2010. *On Social Semantics in Information Retrieval: From Knowledge Discovery to Collective Web Intelligence in the Social Semantic Web*. Ph.D. thesis, University of Bielefeld.