Comparing Corpus-based to Web-based Lookup Techniques for Automatic English Inclusion Detection

Beatrice Alex
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW, Scotland, UK
balex@inf.ed.ac.uk

Abstract

The influence of English as a global language continues to grow to an extent that its words and expressions permeate the original forms of other languages. This paper evaluates a modular Web-based sub-component of an existing English inclusion classifier and compares it to a corpus-based lookup technique. Both approaches are evaluated on a German gold standard data set. It is demonstrated to what extent the Web-based approach benefits from the amount of data available online and the fact that this data is constantly updated.

1 Introduction

The influence of English as a global language continues to grow to an extent that its words and expressions permeate the original forms of many other languages, particularly in domains such as science and technology, commerce, advertising, and current affairs. This is an instance of language mixing driven mainly by globalisation and the growth of the internet, whereby inclusions from other languages appear in an otherwise monolingual text. In this paper, we focus on the automatic detection of English inclusions in German text. Anglicisms and other borrowings from English form by far the most frequent foreign inclusions in German. In specific domains, no less than 6% of German newspaper text can be made up of English tokens (see Section 3).

Previous work reported by Alex (2006; 2005) has focused on devising a classifier that detects anglicisms and other English inclusions in text written in other languages, namely German and French. This English inclusion classifier is based on combining a high precision lexicon module with a high recall search engine module as well as a post-processing step. While the lexicon module allows to classify known English tokens, the search engine module deals well with unknown words, including recent borrowings that have not yet been entered into dictionaries and lexicons. In this paper, we determine the merit of conducting Web-based English inclusion detection using this search engine module compared to using lookup of fixed corpora. We investigate whether the performance of a Web-based approach to English inclusion detection can be achieved by using static corpora and how much data is required to attain a similar performance. We describe two English inclusion detection experiments using a fixed corpus and increasing corpus sub-sets instead of estimated Web corpus sizes. We thereby demonstrate that the search engine module exploits the fact that the amount of data published on the internet is extremely large and that this data is updated continuously.

This paper is organised as follows: after reviewing related work on foreign inclusion detection in Section 2, English inclusions are defined and the English inclusion classifier is described in Section 3. Section 4 presents an overview of the experiments that were conducted as part of this paper. Their results are reported and discussed in Section 5 which is followed by a conclusion in Section 6.

2 Related Work

Conventional language identification systems are successful in recognising the language of larger portions of text but are not well suited to classify individual tokens or subparts thereof. More recently, there have been several initial efforts in devising and evaluating algorithms to identify foreign language portions when the language mixing occurs at the level of the phrase or the word, for example. Pfister and Romsdorfer (2003) proposed to identify foreign inclusions of English and French origin in German text by means of a method that relies on morpho-syntactic analysis combined with lexicon lookup. Marcadet et al. (2005) reported on a combination of different methods including dictionary lookup and character n-gram statistics to detect foreign inclusions in several languages. Furthermore, Farrugia (2005) proposed identifying English code-switching in Maltese text messages using a system that combines Hidden Markov Model language tagging with dictionary lookup and character-based n-gram modelling. Finally, Andersen (2005) tested a series of matching techniques to identify anglicisms in a list of neologisms extracted from Norwegian text whereby a combined chargram and regular expression matching performed best. These previously proposed methods were either not evaluated, not evaluated on unseen data or evaluated on relatively small data sets. However, all of them signal a need for automatically detecting such foreign inclusions for example in the pre-processing of text-to-speech synthesis systems or as an aid to linguists and lexicographers. Other applications and fields which could potentially benefit from such technology are parsing (Alex et al., 2007) and machine translation.

To the best of our knowledge, a Web-based approach to English inclusion detection has not been investigated in previous research. However, in recent years the Web has been exploited by a number of different research initiatives for various natural language processing tasks and entire workshops such as “Web as Corpus” (2006, 2007 and 2008) are now dedicated to this subject area.

Domain	EI tokens	EI types	EI TTR	Accuracy	Precision	Recall	F1
Development set	6.0%	6.8%	0.29	98.07%	93.48%	73.31%	82.17
Unseen test set	6.4%	5.9%	0.25	97.93%	92.13%	75.82%	83.18

Table 1: English inclusion (EI) token and type statistics, EI type-token-ratios (TTR) as well as accuracy, precision, recall and F1-scores for the German internet & telecoms development set and the unseen test set.

3 English Inclusion Detection

The aim of this paper is to evaluate a modular subcomponent of an automatic English inclusion classifier that is designed to identify English forms in other languages. This section first explains what English inclusions are and then describes how they are recognised by the English inclusion classifier that is used for the experiments described in Section 4.

3.1 English Inclusions

The influence of English on other languages has developed into a perennial issue of language contact research which is driven by a mixture of scholarly and public interest in the matter. Onysko's (2007) formal definition of anglicism serves as a framework for this work. He treats the concept anglicism as a hypernym of all English forms occurring in German, including borrowing, code-switching, hybrid forms, pseudo-anglicisms as well as interference and unobtrusive borrowing. Onysko groups the first four categories into core anglicisms and the next two into borderline anglicisms. Essentially, core anglicisms are, with some exceptions that are explained later, the forms that the English inclusion classifier is able to recognise. Interference, i.e. semantic and functional transfer on lexical, semantic, and pragmatic levels as a result of formal similarity of source and receiver language units like *realisieren* (to become aware of)[1], and unobtrusive borrowings like *Keks* (biscuit, from cakes) are not recognised by the classifier as they are formally unmarked in German. Onysko's definition of anglicism is also a preferable theoretic framework for this work as it excludes all semantic borrowing, i.e. loan substitutions (or loan coinage) which are an integral part of other definitions (Betz, 1936; Haugen, 1950; Duckworth, 1977; Carstensen, 1979; Kirkness, 1984; Yang, 1990).
Currently, the English inclusion classifier is designed to recognise but not distinguish between the following types of anglicisms:

- Borrowings: *Business*, *Event*, *Software*
- Code-switching: *real big french guy*, *Gentlemen's Agreement*, *nothing at all*
- English morphemes in hyphenated hybrid forms: *Airline*-Aktien (airline share), *Computer*-Maus (computer mouse), *Online*-Dienst (online service)
- Pseudo-anglicisms: *Beamer* (video projector), *Handy* (mobile phone), *Oldtimer* (vintage car)

[1]In German, the verb *realisieren* used only to be used in the sense of *to carry out*, or *to put into practice*. Because of its similarity to the English verb *realise*, it has adopted a new sense, as in *to become aware of sth.*

Two important linguistic processes in German are compounding and inflection which need to be considered as English inclusions are also affected by them. Both phenomena result in the formation of hybrid, or mixed-lingual forms, in this case specifically tokens made up of English and German morphemes. The English inclusion classifier described in the following section is currently only designed to recognise English morphemes in hyphenated hybrid forms. In future work, the aim is to extend the classifier to recognise English inclusions in morphological derived forms and in other types of hybrid forms as well.

3.2 English Inclusion Classifier

The English inclusion classifier used for the experiments described in this paper consists of three classification modules: a lexicon and a search engine module as well as a post-processing step. The lexicon lookup is performed using the German and English CELEX lexicons to classify all known words. Tokens only found in the English lexicon are classified as English. Tokens found in both databases are classified by the post-processing module. Unknown tokens, i.e. tokens found in neither lexicon are passed to the search engine module. The latter performs language classification based on the maximum normalised score of the number of hits returned for two searches per token, one for each language (Alex, 2005). This score is determined by weighting the number of hits, i.e. the "absolute frequency" by the estimated size of the accessible Web corpus for that language. This Web corpus estimation is motivated by Grefenstette and Niochi (2000). In the following section, the performance of this search engine module, which has access to extremely large quantities of data on the Web, is compared against a corpus search module where this access is limited to a fixed corpus. The English inclusion classifier's final component, the rule-based post-processing module, classifies single-character tokens and resolves language classification ambiguities for interlingual homographs, English function words, names of currencies and units of measurement. A further post-processing step relates language information between abbreviations or acronyms and their definitions in combination with an abbreviation extraction algorithm (Schwartz and Hearst, 2003). Finally, several rules disambiguate English inclusions from person names (Alex, 2006).

For German and French, the classifier has been evaluated on unseen test sets in different domains, including internet & telecoms, space travel and European Union related topics (Alex, 2006). Table 1 presents an overview of the percentages of English inclusion tokens and types within the gold standard annotation of the German development and test sets for the internet & telecoms domain, and illustrates

how well the English inclusion classifier is able to detect them in terms of F1-score.[2] The figures show that the frequencies of English inclusions are similar in both sets, with slightly more repetition in the test set, and that the classifier is able to detect them equally well with an F1-score of over 82 point for each data set.[3]

4 Experiments

In order to understand the merit of the search engine module and the amount of data it can access better, in the following experiments, the search engine module is replaced with a corpus search module that determines relative token frequencies based on fixed corpora. Here, the language classification is essentially based on real corpus frequencies rather than estimated Web corpus frequencies. Language identification is simply conducted as a result of the higher relative frequency (rf) of a token (t) for a given corpus (C) in a particular language (L) and calculated as the actual frequency of a token in the corpus normalised by the corpus size (N).

$$rf_{C(L)}(t) = \frac{f_{C(L)}(t)}{N_{C(L)}} \quad (1)$$

If the relative frequency of the token in the English corpus is higher than that in the corpus of the base language of the text, the token is classed as English. This experimental setup therefore requires two corpora, one for the inclusion language (English) and one for the base language of the text (German). In the first experiment, two corpora of roughly equal size were used: the Wall Street Journal section of the Penn Treebank corpus, Version 3.0 (Marcus et al., 1993) amounting to around 1.2m tokens and the combined German NEGRA and TIGER corpora (Skut et al., 1998; Brants et al., 2002) containing approximately 1.1m tokens. Both data sets were published in the 1990s. For the purpose of determining the relative frequencies of a given token for both languages and identifying its language accordingly, the corpora were converted into frequency lists.
All subsequent corpus search experiments are conducted using the German development set of newspaper articles in the internet & telecoms domain, a set containing a relatively high percentage of English inclusions. The architecture of the classifier is essentially the same as that of the English inclusion classifier, except that the search engine module is replaced by the corpus search module. Relative token frequencies are calculated using the same equations as in the search engine module, but based on a fixed corpus, instead of an estimated Web corpus for each language. The corpus search engine module is preceded by the pre-processing and lexicon modules and followed by optional post-processing.
A second experiment was conducted to test the hypothesis that the search engine module performs better due to the large amount of data it can access, and the fact that this data is constantly updated and enriched with new material. The aim is to simulate the search engine module's behaviour in a more controlled fashion by making use of increasing corpus sub-sets. These are drawn from a corpus more recently released than the Wall Street Journal corpus, the Agence France Presse content of the English Gigaword corpus[4] (published between 1994-1997 and 2001-2002). The corpus sub-sets are created by randomly selecting sentences from the Gigaword corpus amounting to 1m, 10m, 20m, 30m and 40m tokens. While the German corpus (combined NEGRA/TIGER) remains unchanged, each of the English corpus sub-sets are used by the corpus search module in a separate run of the classifier over the German internet & telecoms development data. The idea is to grant the corpus search module access to more and more data in order to identify the language of individual tokens.

5 Results and Discussion

As can be seen in Table 2 (Experiment 1), using the Wall Street Journal corpus as the basis for token-level language identification in the corpus search module only increases the performance of the English inclusion classifier by 9.36 to 46.10 points in F1-score compared to running the lexicon module alone. This score is much lower than the performance achieved with the combined lexicon and search engine module (76.58). The relatively poor result of the corpus search module is partially caused by the fact that the English Wall Street Journal corpus is limited in size and may therefore not cover the English terms that occur in the articles belonging to the German development set. Conversely, the likelihood that a word is not found online is very small given that search engines have access to billions of words. The other reason for the low score is the time period during which the text in the Wall Street Journal corpus was published (1993-1994). While this English corpus is a relatively old collection, the German internet newspaper articles were published more recently between 2001 and 2005. It is therefore extremely likely that the English inclusions, which to some extent are recently emerged technological and computing vocabulary, did not exist or were not commonly used in the early 1990s. Moreover, unlike the German development set, the Wall Street Journal corpus contains general newspaper text not limited to a specific topic. This discrepancy in domain is another crucial factor in the small performance increase of combining the corpus search module with the lexicon module.
Table 3 (Experiment 2) shows the results of making increasingly larger data sets available to the corpus search module. The amount of tokens extracted from the English Gigaword corpus, used in this experiment, is increased incrementally from 1m up to 40m tokens. Results are averages over 5 runs for different selections of increasing corpus sizes and are listed with and without post-processing. As anticipated, granting the corpus search module access to larger amounts of data results in an incremental performance increase in F1. Using an English corpus of 1m tokens, the corpus search module results in an F1-score of

[2] F1-scores refer to the English tokens and are calculated giving equal weight to precision (P) and recall (R) as:
$F1 = (2 * P * R)/(P + R)$.

[3] Inter-annotator agreement for marking up English inclusions in German text was found to be very high at a pairwise F1-score of 91.04 and a κ-score of 0.9075.

[4] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05`

Corpus Size	No. of Types	F1-score without PP	F1-score with PP
Lexicon module only			
N/A		36.74	39.11
Lexicon + corpus search module: Wall Street Journal corpus			
1,173,747	43,808	46.10	48.64
Lexicon + search engine module			
638.9bn tokens (estimate)		76.58	82.17

Table 2: Evaluation of the corpus search module using the Wall Street Journal corpus and the combined NEGRA/TIGER corpus with/without post-processing (PP) compared to the lexicon module only and a combined lexicon and search engine module approach.[5]

Corpus Size	Avg No. of Types	F1-score without PP	F1-score with PP
Lexicon module only			
N/A		36.74	39.11
Lexicon + corpus search module: Gigaword corpus			
1,000,000	52,268	60.37	67.06
10,000,000	165,445	65.41	71.92
20,000,000	229,139	66.73	73.18
30,000,000	273,139	69.74	74.74
40,000,000	308,421	70.89	75.87
Lexicon + search engine module			
638.9bn tokens (estimate)[1]		76.58	82.17

Table 3: Evaluation of the corpus search module with increasing sub-sets of the Gigaword corpus with/without post-processing (PP) compared to the lexicon module only and a combined lexicon and search engine module approach.

60.37, that is 23.63 points higher than when just applying the lexicon module, but 16.21 points lower than when using the search engine module in its place. The fact that this F1-score is 14.27 points higher compared to the one obtained when using the Wall Street Journal (almost equal in size) demonstrates that data currentness is vital for English inclusion detection. The classifier improves steadily with access to larger corpus frequency lists and reaches an F1-score of 70.89 when the corpus search module determines relative token frequencies in an English corpus containing 40m tokens. Figure 1 shows that the performance increases are reduced with larger corpus sizes. However, in order to achieve a similar performance as the search engine module (82.17 and 76.58 with/without post-processing, respectively), the corpus search module would need to have access to much larger data sets.

6 Conclusion

To summarise, it was shown that token-level language identification improves with access to larger data sets. It also emerged that the time of publishing is an important aspect that needs to be considered. The use of any fixed-size corpus for language identification purposes clearly has its drawbacks. Such a collection is unlikely to contain all possible lexical items and, with languages evolving constantly, is out-of-date as soon as it is created and made available. Search engines provide access to extremely large collections of data which are constantly updated and changing with time and language use. Therefore, the search engine module has a clear superiority over accessing a corpus that is a data snap-shot of a particular time period and is limited in size. This is clearly reflected in the performance comparison of both methods. Access to a corpus considerably larger than 40m tokens would be required for the corpus search module to reach the same level of performance as that of the search engine module. This is not necessarily a surprising conclusion. However, testing the corpus-based lookup approach was still justified in order to determine whether it presents a potential alternative to the Web-based English inclusion detection approach, considering that the later is more computationally costly as well as time-consuming and also is limited to the number of searches allowed per day by the underlying search engine.

7 Acknowledgements

I would like to thank Claire Grover and Frank Keller for their comments on this work. This research is supported by grants from the Scottish Enterprise Edinburgh-Stanford Link (R36759) and ESRC as well as the University of Edinburgh.

8 References

Beatrice Alex, Amit Dubey, and Frank Keller. 2007. Using foreign inclusion detection to improve parsing performance. In *Proceedings of EMNLP-CoNLL 2007*, pages 151–160, Prague, Czech Republic.

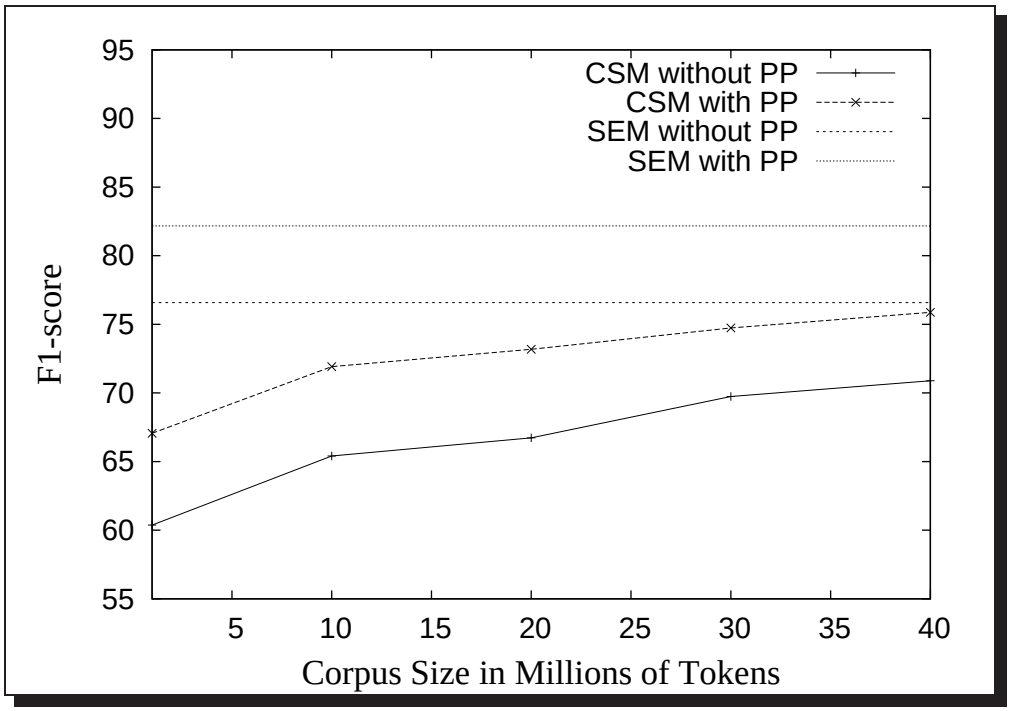

Figure 1: Performance gains using the corpus search module (CSM) with/without post-processing (PP) with increasing sub-sets of the Gigaword corpus, compared to the performance of the search engine module (SEM) represented by the horizontal lines.

Beatrice Alex. 2005. An unsupervised system for identifying English inclusions in German text. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), Student Research Workshop*, pages 133–138, Ann Arbor, Michigan, USA.

Beatrice Alex. 2006. Integrating language knowledge resources to extend the English inclusion classifier to a new language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 595–600, Genoa, Italy.

Gisle Andersen. 2005. Assessing algorithms for automatic extraction of anglicisms in Norwegian texts. In *Proceedings of the International Conference on Corpus Linguistics (CL2005)*, Birmingham, UK. Available online at: `http://www.corpus.bham.ac.uk/PCLC/`.

Werner Betz. 1936. *Der Einfluß des Lateinischen auf den althochdeutschen Sprachschatz*. Winter, Heidelberg.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*, pages 24–41, Sozopol, Bulgaria.

Broder Carstensen. 1979. Evidente und latente Einflüsse des Englischen auf das Deutsche. In Peter Braun, editor, *Fremdwort-Diskussion*, pages 90–94. Fink, Munich.

David Duckworth. 1977. Zur terminologischen und systematischen Grundlage der Forschung auf dem Gebiet der englisch-deutschen Interferenz. In Herbert Kolb and Hartmut Lauffer, editors, *Sprachliche Interferenz*, pages 36–56. Niemeyer, Tübingen.

Paulseph-John Farrugia. 2005. *Text to Speech Technologies for Mobile Telephony Services*. Department of Computer Science and AI, University of Malta, Msida, Malta. MSc Thesis.

Gregory Grefenstette and Julien Nioche. 2000. Estimation of English and non-English language use on the WWW. In *Proceedings of RIAO (Recherche d'Informations Assistée par Ordinateur) 2000*, pages 237–246, Paris, France.

Einar Haugen. 1950. The analysis of linguistic borrowing. *Language*, 26:210–231.

Alan Kirkness. 1984. Aliens, denziens, hybrids and natives: foreign influence on the etymological structure of German vocabulary. In Charles V. J. Russ, editor, *Foreign Influences on German*. Lochee, Dundee.

J. C. Marcadet, V. Fischer, and C. Waast-Richard. 2005. A transformation-based learning approach to language identification for mixed-lingual text-to-speech synthesis. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005 - Eurospeech)*, pages 2249–2252, Lisbon, Portugal.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Alexander Onysko. 2007. *Anglicisms in German: Borrowing, Lexical Productivity and Written Codeswitching*. De Gruyter, Berlin/New York.

Beat Pfister and Harald Romsdorfer. 2003. Mixed-lingual analysis for polyglot TTS synthesis. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2037–2040, Geneva, Switzerland.

Ariel Schwartz and Marti Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing (PSB 2003)*, pages 451–462, Kauai, Hawaii.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *Proceedings of the Conference on Language Resources and Evaluation (LREC 1998)*, pages 705–712, Granada, Spain.

Wenliang Yang. 1990. *Anglizismen im Deutschen: am Beispiel des Nachrichtenmagazins Der Spiegel*. Niemeyer, Tübingen.