An Unsupervised System for Identifying English Inclusions in German Text

Beatrice Alex
School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, UK
v1balex@inf.ed.ac.uk

Abstract

We present an unsupervised system that exploits linguistic knowledge resources, namely English and German lexical databases and the World Wide Web, to identify English inclusions in German text. We describe experiments with this system and the corpus which was developed for this task. We report the classification results of our system and compare them to the performance of a trained machine learner in a series of in- and cross-domain experiments.

1 Introduction

The recognition of foreign words and foreign named entities (NEs) in otherwise mono-lingual text is beyond the capability of many existing approaches and is only starting to be addressed. This language mixing phenomenon is prevalent in German where the number of anglicisms has increased considerably.

We have developed an unsupervised and highly efficient system that identifies English inclusions in German text by means of a computationally inexpensive lookup procedure. By unsupervised we mean that the system does not require any annotated training data and only relies on lexicons and the Web. Our system allows linguists and lexicographers to observe language changes over time, and to investigate the use and frequency of foreign words in a given language and domain. The output also represents valuable information for a number of applications, including polyglot text-to-speech (TTS) synthesis and machine translation (MT).

We will first explain the issue of foreign inclusions in German text in greater detail with examples in Section 2. Sections 3 and 4 describe the data we used and the architecture of our system. In Section 5, we provide an evaluation of the system output and compare the results with those of a series of in- and cross-domain machine learning experiments outlined in Section 6. We conclude and outline future work in Section 7.

2 Motivation

In natural language, new inclusions typically fall into two major categories, foreign words and proper nouns. They cause substantial problems for NLP applications because they are hard to process and infinite in number. It is difficult to predict which foreign words will enter a language, let alone create an exhaustive gazetteer of them. In German, there is frequent exposure to documents containing English expressions in business, science and technology, advertising and other sectors. A look at current headlines confirms the existence of this phenomenon:

(1) "*Security-Tool* verhindert, dass *Hacker* über *Google* Sicherheitslücken finden"[1]
Security tool prevents hackers from finding security holes via Google.

An automatic classifier of foreign inclusions would prove valuable for linguists and lexicographers who

[1] Published in Computerwelt on 10/01/2005: http://www.computerwelt.at

Proceedings of the ACL Student Research Workshop, pages 133–138,
Ann Arbor, Michigan, June 2005. ©2005 Association for Computational Linguistics

study this language-mixing phenomenon because lexical resources need to be updated and reflect this trend. As foreign inclusions carry critical content in terms of pronunciation and semantics, their correct recognition will also provide vital knowledge in applications such as polyglot TTS synthesis or MT.

3 Data

Our corpus is made up of a random selection of online German newspaper articles published in the Frankfurter Allgemeine Zeitung between 2001 and 2004 in the domains of (1) *internet & telecomms*, (2) *space travel* and (3) *European Union*. These domains were chosen to examine the different use and frequency of English inclusions in German texts of a more technological, scientific and political nature. With approximately 16,000 tokens per domain, the overall corpus comprises of 48,000 tokens (Table 1).

We created a manually annotated gold standard using an annotation tool based on NITE XML (Carletta et al., 2003). We annotated two classes whereby English words and abbreviations that expand to English terms were classed as "English" (EN) and all other tokens as "Outside" (O).[2] Table 1 presents the number of English inclusions annotated in each gold standard set and illustrates that English inclusions are very sparse in the EU domain (49 tokens) but considerably frequent in the documents in the internet and space travel domains (963 and 485 tokens, respectively). The type-token ratio (TTR) signals that the English inclusions in the space travel data are less diverse than those in the internet data.

Domain		Tokens	Types	TTR
Internet	Total	15919	4152	0.26
	English	963	283	0.29
Space	Total	16066	3938	0.25
	English	485	73	0.15
EU	Total	16028	4048	0.25
	English	49	30	0.61

Table 1: English token and type statistics and type-token-ratios (TTR) in the gold standard

4 System Description

Our system is a UNIX pipeline which converts HTML documents to XML and applies a set of modules to add linguistic markup and to classify nouns as German or English. The pipeline is composed of a pre-processing module for tokenisation and POS-tagging as well as a lexicon lookup and Google lookup module for identifying English inclusions.

4.1 Pre-processing Module

In the pre-processing module, the downloaded Web documents are firstly cleaned up using `Tidy`[3] to remove HTML markup and any non-textual information and then converted into XML. Subsequently, two rule-based grammars which we developed specifically for German are used to tokenise the XML documents. The grammar rules are applied with `lxtransduce`[4], a transducer which adds or rewrites XML markup on the basis of the rules provided. `Lxtransduce` is an updated version of `fsgmatch`, the core program of LT TTT (Grover et al., 2000). The tokenised text is then POS-tagged using TnT trained on the German newspaper corpus Negra (Brants, 2000).

4.2 Lexicon Lookup Module

For the initial lookup, we used CELEX, a lexical database of English, German and Dutch containing full and inflected word forms as well as corresponding lemmas. CELEX lookup was only performed for tokens which TnT tagged as nouns (NN), foreign material (FM) or named entities (NE) since anglicisms representing other parts of speech are relatively infrequent in German (Yeandle, 2001). Tokens were looked up twice, in the German and the English database and parts of hyphenated compounds were checked individually. To identify capitalised English tokens, the lookup in the English database was made case-insensitive. We also made the lexicon lookup sensitive to POS tags to reduce classification errors. Tokens were found either only in the German lexicon (1), only in the English lexicon (2) in both (3) or in neither lexicon (4).

(1) The majority of tokens found exclusively in

[2] We did not annotate English inclusions if part of URLs (*www.stepstone.de*), mixed-lingual unhyphenated compounds (*Shuttleflug*) or with German inflections (*Receivern*) as further morphological analysis is required to recognise them. Our aim is to address these issues in future work.

[3] `http://tidy.sourceforge.net`

[4] `http://www.ltg.ed.ac.uk/~richard/lxtransduce.html`

the German lexicon are actual German words. The remaining are English words with German case inflection such as *Computern*. The word *Computer* is used so frequently in German that it already appears in lexicons and dictionaries. To detect the base language of the latter, a second lookup can be performed checking whether the lemma of the token also occurs in the English lexicon.

(2) Tokens found exclusively in the English lexicon such as *Software* or *News* are generally English words and do not overlap with German lexicon entries. These tokens are clear instances of foreign inclusions and consequently tagged as English.

(3) Tokens which are found in both lexicons are words with the same orthographic characteristics in both languages. These are words without inflectional endings or words ending in *s* signalling either the German genitive singular or the German and English plural forms of that token, e.g. *Computers*. The majority of these lexical items have the same or similar semantics in both languages and represent assimilated loans and cognates where the language origin is not always immediately apparent. Only a small subgroup of them are clearly English loan words (e.g. *Monster*). Some tokens found in both lexicons are interlingual homographs with different semantics in the two languages, e.g. *Rat* (*council* vs. *rat*). Deeper semantic analysis is required to classify the language of such homographs which we tagged as German by default.

(4) All tokens found in neither lexicon are submitted to the Google lookup module.

4.3 Google Lookup Module

The Google lookup module exploits the World Wide Web, a continuously expanding resource with documents in a multiplicity of languages. Although the bulk of information available on the Web is in English, the number of texts written in languages other than English has increased rapidly in recent years (Crystal, 2001; Grefenstette and Nioche, 2000).

The exploitation of the Web as a linguistic corpus is developing into a growing trend in computational linguistics. The sheer size of the Web and the continuous addition of new material in different languages make it a valuable pool of information in terms of language in use. The Web has already been used successfully for a series of NLP tasks such as MT (Grefenstette, 1999), word sense disambiguation (Agirre and Martinez, 2000), synonym recognition (Turney, 2001), anaphora resolution (Modjeska et al., 2003) and determining frequencies for unseen bi-grams (Keller and Lapata, 2003).

The Google lookup module obtains the number of hits for two searches per token, one on German Web pages and one on English ones, an advanced language preference offered by Google. Each token is classified as either German or English based on the search that returns the higher normalised score of the number of hits. This score is determined by weighting the number of raw hits by the size of the Web corpus for that language. We determine the latter following a method proposed by Grefenstette and Niochi (2000) by using the frequencies of a series of representative tokens within a standard corpus in a language to determine the size of the Web corpus for that language. We assume that a German word is more frequently used in German text than in English and vice versa. As illustrated in Table 2, the German word *Anbieter* (provider) has a considerably higher weighted frequency in German Web documents (DE). Conversely, the English word *provider* occurs more often in English Web documents (EN). If both searches return zero hits, the token is classified as German by default. Word queries that return zero or a low number of hits can also be indicative of new expressions that have entered a language.

Google lookup was only performed for the tokens found in neither lexicon in order to keep computational cost to a minimum. Moreover, a preliminary experiment showed that the lexicon lookup is already sufficiently accurate for tokens contained exclusively in the German or English databases. Current Google search options are also limited in that queries cannot be treated case- or POS-sensitively. Consequently, interlingual homographs would often mistakenly be classified as English.

Language	DE		EN	
Hits	Raw	Normalised	Raw	Normalised
Anbieter	3.05	0.002398	0.04	0.000014
Provider	0.98	0.000760	6.42	0.002284

Table 2: Raw counts (in million) and normalised counts of two Google lookup examples

5 Evaluation of the Lookup System

We evaluated the system's performance for all tokens against the gold standard. While the accuracies in Table 3 represent the percentage of all correctly tagged tokens, the F-scores refer to the English tokens and are calculated giving equal weight to precision (P) and recall (R) as $F = (2 * P * R)/(P + R)$.

The system yields relatively high F-scores of 72.4 and 73.1 for the internet and space travel data but only a low F-score of 38.6 for the EU data. The latter is due to the sparseness of English inclusions in that domain (Table 1). Although recall for this data is comparable to that of the other two domains, the number of false positives is high, causing low precision and F-score. As the system does not look up one-character tokens, we implemented further post-processing to classify individual characters as English if followed by a hyphen and an English inclusion. This improves the F-score by 4.8 for the internet data to 77.2 and by 0.6 for the space travel data to 73.7 as both data sets contain words like *E-Mail* or *E-Business*. Post-processing does not decrease the EU score. This indicates that domain-specific post-processing can improve performance.

Baseline accuracies when assuming that all tokens are German are also listed in Table 3. As F-scores are calculated based on the English tokens in the gold standard, we cannot report comparable baseline F-scores. Unsurprisingly, the baseline accuracies are relatively high as most tokens in a German text are German and the amount of foreign material is relatively small. The added classification of English inclusions yielded highly statistical significant improvements ($p<0.001$) over the baseline of 3.5% for the internet data and 1.5% for the space travel data. When classifying English inclusions in the EU data, accuracy decreased slightly by 0.3%.

Table 3 also shows the performance of `TextCat`, an n-gram-based text categorisation algorithm of Cavnar and Trenkle (1994). While this language idenfication tool requires no lexicons, its F-scores are low for all 3 domains and very poor for the EU data. This confirms that the identification of English inclusions is more difficult for this domain, coinciding with the result of the lookup system. The low scores also prove that such language identification is unsuitable for token-based language classification.

Domain	Method	Accuracy	F-score
Internet	Baseline	94.0%	-
	Lookup	97.1%	72.4
	Lookup + post	97.5%	77.2
	TextCat	92.2%	31.0
Space	Baseline	97.0%	-
	Lookup	98.5%	73.1
	Lookup + post	98.5%	73.7
	TextCat	93.8%	26.7
EU	Baseline	99.7%	-
	Lookup	99.4%	38.6
	Lookup + post	99.4%	38.6
	TextCat	96.4%	4.7

Table 3: Lookup results (with and without post-processing) compared to TextCat and baseline

6 Machine Learning Experiments

The recognition of foreign inclusions bears great similarity to classification tasks such as named entity recognition (NER), for which various machine learning techniques have proved successful. We were therefore interested in determining the performance of a trained classifier for our task. We experimented with a conditional Markov model tagger that performed well on language-independent NER (Klein et al., 2003) and the identification of gene and protein names (Finkel et al., 2005).

6.1 In-domain Experiments

We performed several 10-fold cross-validation experiments with different feature sets. They are referred to as in-domain (ID) experiments as the tagger is trained and tested on data from the same domain (Table 4). In the first experiment (ID1), we use the tagger's standard feature set including words, character sub-strings, word shapes, POS-tags, abbreviations and NE tags (Finkel et al., 2005). The resulting F-scores are high for the internet and space travel data (84.3 and 91.4) but are extremely low for the EU data (13.3) due to the sparseness of English inclusions in that data set. ID2 involves the same setup as ID1 but eliminating all features relying on the POS-tags. The tagger performs similarly well for the internet and space travel data but improves by 8 points to an F-score of 21.3 for the EU data. This can be attributed to the fact that the POS-tagger does not perform with perfect accuracy particularly on data containing foreign inclusions. Providing the tagger with this information is therefore not necessarily useful for this task, especially when the data is sparse. Nevertheless, there is a big discrepancy between the F-score for the EU data and those of the other two data sets. ID3 and ID4 are set up as ID1 and ID2 but incorporating the output of the lookup system as a gazetteer feature. The tagger benefits considerably from this lookup feature and yields better F-scores for all three domains in ID3 (internet: 90.6, space travel: 93.7, EU: 44.4).

Table 4 also compares the best F-scores produced with the tagger's own feature set (ID2) to the best results of the lookup system and the baseline. While the tagger performs much better for the internet and the space travel data, it requires hand-annotated training data. The lookup system, on the other hand, is essentially unsupervised and therefore much more portable to new domains. Given the necessary lexicons, it can easily be run over new text and text in a different language or domain without further cost.

6.2 Cross-domain Experiments

The tagger achieved surprisingly high F-scores for the internet and space travel data, considering the small training data set of around 700 sentences used for each ID experiment described above. Although both domains contain a large number of English inclusions, their type-token ratio amounts to 0.29 in the internet data and 0.15 in the space travel data (Table 1), signalling that English inclusions are frequently repeated in both domains. As a result, the likelihood of the tagger encountering an unknown inclusion in the test data is relatively small.

To examine the tagger's performance on a new domain containing more unknown inclusions, we ran two cross-domain (CD) experiments: CD1, training on the internet and testing on the space travel data, and CD2, training on the space travel and testing on the internet data. We chose these two domain pairs to ensure that both the training and test data contain a relatively large number of English inclusions. Table 5 shows that the F-scores for both CD experiments are much lower than those obtained when training and testing the tagger on documents from the same domain. In experiment CD1, the F-score only amounts to 54.2 while the percentage of unknown target types in the space travel test data is 81.9%. The F-score is even lower in the second experiment at 22.2 which can be attributed to the fact that the percentage of unknown target types in the internet test data is higher still at 93.9%.

Domain		Accuracy	F-score
Internet	ID1	98.4%	84.3
	ID2	98.3%	84.3
	ID3	98.9%	90.6
	ID4	98.9%	90.8
	Best Lookup	97.5%	77.2
	Baseline	94.0%	-
Space	ID1	99.5%	91.4
	ID2	99.5%	91.3
	ID3	99.6%	93.7
	ID4	99.6%	92.8
	Best Lookup	98.5%	73.7
	Baseline	97.0%	-
EU	ID1	99.7%	13.3
	ID2	99.7%	21.3
	ID3	99.8%	44.4
	ID4	99.8%	44.4
	Best Lookup	99.4%	38.6
	Baseline	99.7%	-

Table 4: Accuracies and F-scores for ID experiments

	Accuracy	F-score	UTT
CD1	97.9%	54.2	81.9%
Best Lookup	98.5%	73.7	-
Baseline	97.0%	-	-
CD2	94.6%	22.2	93.9%
Best Lookup	97.5%	77.2	-
Baseline	94.0%	-	-

Table 5: Accuracies, F-scores and percentages of unknown target types (UTT) for cross-domain experiments compared to best lookup and baseline

These results indicate that the tagger's high performance in the ID experiments is largely due to the fact that the English inclusions in the test data are known, i.e. the tagger learns a lexicon. It is therefore more complex to train a machine learning classifier to perform well on new data with more and more new anglicisms entering German over time. The amount of unknown tokens will increase constantly unless new annotated training data is added.

7 Conclusions and Future Work

We have presented an unsupervised system that exploits linguistic knowledge resources including lexicons and the Web to classify English inclusions in German text on different domains. Our system can be applied to new texts and domains with little computational cost and extended to new languages as long as lexical resources are available. Its main advantage is that no annotated training data is required.

The evaluation showed that our system performs well on non-sparse data sets. While being outperformed by a machine learner which requires a trained model and therefore manually annotated data, the output of our system increases the performance of the learner when incorporating this information as an additional feature. Combining statistical approaches with methods that use linguistic knowledge resources can therefore be advantageous.

The low results obtained in the CD experiments indicate however that the machine learner merely learns a lexicon of the English inclusions encountered in the training data and is unable to classify many unknown inclusions in the test data. The Google lookup module implemented in our system represents a first attempt to overcome this problem as the information on the Web never remains static and at least to some extent reflects language in use.

The current system tracks full English word forms. In future work, we aim to extend it to identify English inclusions within mixed-lingual tokens. These are words containing morphemes from different languages, e.g. English words with German inflection (*Receivern*) or mixed-lingual compounds (*Shuttleflug*). We will also test the hypothesis that automatic classification of English inclusions can improve text-to-speech synthesis quality.

Acknowledgements

Thanks go to Claire Grover and Frank Keller for their input. This research is supported by grants from the University of Edinburgh, Scottish Enterprise Edinburgh-Stanford Link (R36759) and ESRC.

References

Eneko Agirre and David Martinez. 2000. Exploring automatic word sense disambiguation with decision lists and the Web. In *Proceedings of the Semantic Annotation and Intelligent Annotation workshop, COLING*.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*.

Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holgar Voormann. 2003. The NITE XML toolkit: flexible annotation for multimodal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press.

Jenny Finkel, Shipra Dingare, Christopher Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. 2005. Exploring the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics*. In press.

Gregory Grefenstette and Julien Nioche. 2000. Estimation of English and non-English language use on the WWW. In *Proceedings of RIAO 2000*.

Gregory Grefenstette. 1999. The WWW as a resource for example-based machine translation tasks. In *Proceedings of ASLIB'99 Translating and the Computer*.

Claire Grover, Colin Matheson, Andrei Mikheev, and Moens Marc. 2000. LT TTT - a flexible tokenisation tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Frank Keller and Mirella Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):458–484.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the 7th Conference on Natural Language Learning*.

Natalia Modjeska, Katja Markert, and Malvina Nissim. 2003. Using the Web in machine learning for other-anaphora resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Peter D. Turney. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*.

David Yeandle. 2001. Types of borrowing of Anglo-American computing terminology in German. In Marie C. Davies, John L. Flood, and David N. Yeandle, editors, *Proper Words in Proper Places: Studies in Lexicology and Lexicography in Honour of William Jervis Jones*, pages 334–360. Stuttgarter Arbeiten zur Germanistik 400, Stuttgart, Germany.