Bridging Languages through Etymology:
The case of cross language text categorization

Vivi Nastase and **Carlo Strapparava**
Human Language Technologies, Fondazione Bruno Kessler
Trento, Italy
{nastase, strappa}@fbk.eu

Abstract

We propose the hypothesis that word etymology is useful for NLP applications as a bridge between languages. We support this hypothesis with experiments in cross-language (English-Italian) document categorization. In a straightforward bag-of-words experimental set-up we add etymological ancestors of the words in the documents, and investigate the performance of a model built on English data, on Italian test data (and viceversa). The results show not only statistically significant, but a large improvement – a jump of almost 40 points in F1-score – over the raw (vanilla bag-of-words) representation.

1 Introduction

When exposed to a document in a language he does not know, a reader might be able to glean some meaning from words that are the same (e.g. names) or similar to those in a language he knows. As an example, let us say that an Italian speaker is reading an English text that contains the word *expense*, which he does not know. He may be reminded however of the Latin word *expensa* which is also the etymological root of the Italian word *spesa*, which usually means "cost"/"shopping", and may thus infer that the English word refers to the cost of things. In the experiments presented here we investigate whether an automatic text categorization system could benefit from knowledge about the etymological roots of words. The cross language text categorization (CLTC) task consists of categorizing documents in a target language L_t using a model built from labeled examples in a source language L_s. The task becomes more difficult when the data consists of comparable corpora in the two languages – documents on the same topics (e.g. sports, economy) – instead of parallel corpora – there exists a one-to-one correspondence between documents in the corpora for the two languages, one document being the translation of the other.

To test the usefulness of etymological information we work with comparable collections of news articles in English and Italian, whose articles are assigned one of four categories: *culture_and_school*, *tourism*, *quality_of_life*, *made_in_Italy*. We perform a progression of experiments, which embed etymological information deeper and deeper into the model. We start with the basic set-up, representing the documents as bag-of-words, where we train a model on the English training data, and use this model to categorize documents from the Italian test data (and viceversa). The results are better than random, but quite low. We then add the etymological roots of the words in the data to the bag-of-words, and notice a large – 21 points – increase in performance in terms of F1-score. We then use the bag-of-words representation of the training data to build a semantic space using LSA, and use the generated word vectors to represent the training and test data. The improvement is an additional 16 points in F1-score.

Compared to related work, presented in Section 3, where cross language text categorization is approached through translation or mapping of features (i.e. words) from the source to the target language, word etymologies are a novel source of cross-lingual knowledge. Instead of mapping features between languages, we introduce new features which are shared, and thus do not need translation or other forms of mapping.

The experiments presented show unequivocally that word etymology is a useful addition to computational models, just as they are to readers who have such knowledge. This is an interesting and useful result, especially in the current research landscape where using and exploiting multi-linguality is a desired requirement.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 651–659,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

morpheme	relation	related morpheme
eng: ex-	rel:etymological_origin_of	eng: excentric
eng: expense	rel:etymology	lat: expensa
eng: -ly	rel:etymological_origin_of	eng: absurdly
eng: -ly	rel:etymological_origin_of	eng: admirably
...		
ita: spesa	rel:etymology	lat: expensa
ita: spesa	rel:has_derived_form	ita: spese
...		
ita: spesare	rel:etymologically_related	ita: spesa
...		
lat: expensa	rel:etymological_origin_of	eng: expense
lat: expensa	rel:etymological_origin_of	ita: spesa
...		
lat: expensa	rel:is_derived_from	lat: expensus
...		

English: *muscle*
↓
French: *muscle*
↓
Latin: *musculus*
↓
Latin: *mus*
↓
Proto Indo-European: *muh*$_2$*s*

Figure 1: Sample entries from the Etymological WordNet, and a few etymological layers

2 Word Etymology

Word etymology gives us a glimpse into the evolution of words in a language. Words may be adopted from a language because of cultural, scientific, economic, political or other reasons (Hitchings, 2009). In time these words "adjust" to the language that adopted them – their sense may change to various degrees – but they are still semantically related to their etymological roots. To illustrate the point, we show an example that the reader, too, may find amusing: on the ticket validation machine on Italian buses, by way of instruction, it is written *Per obliterare il biglietto* .... A native/frequent English speaker would most probably key in on, and be puzzled by, the word *obliterare*, very similar to the English *obliterate*, whose most used sense is *to destroy completely / cause to physically disappear* . The Italian *obliterare* has the "milder" sense of *cancellare* – *cancel* (which is also shared by the English *obliterate*, but is less frequent according to Merriam-Webster), and both come from the Latin *obliterare* – erase, efface, cause to disappear. While there has been some sense migration – in English the more (physically) destructive sense of the word has higher prominence, while in Italian the word is closer in meaning to its etymological root – the Italian and the English words are still semantically related.

Dictionaries customarily include etymological information for their entries, and recently, Wikipedia's Wiktionary has joined this trend. The etymological information can, and indeed has been extracted and prepared for machine consumption (de Melo and Weikum, 2010): Etymological WordNet[1] contains 6,031,431 entries for 2,877,036 words (actually, morphemes) in 397 languages. A few sample entries from this resource are shown in Figure 1.

The information in Etymological WordNet is organized around 5 relations: *etymology* with its inverse *etymological_origin_of*; *is_derived_from* with its inverse *has_derived_form*; and the symmetrical *etymologically_related*. The *etymology* relation links a word with its etymological ancestors, and it is the relation used in the experiments presented here. Prefixes and suffixes – such as *ex-* and *-ly* shown in Figure 1 – are filtered out, as they bring in much noise by relating words that merely share such a morpheme (e.g. *absurdly* and *admirably*) but are otherwise semantically distant. *has_derived_form* is also used, to capture morphological variations.

The depth of the etymological hierarchy (considering the *etymology* relations) is 10. Figure 1 shows an example of a word with several levels of etymological ancestry.

[1] http://www1.icsi.berkeley.edu/~demelo/etymwn/

$$\left[\begin{array}{ll|ccccc|ccccc}
 & & \multicolumn{5}{c|}{\textit{English texts}} & \multicolumn{5}{c}{\textit{Italian texts}} \\
 & & t_1^e & t_2^e & \cdots & t_{n-1}^e & t_n^e & t_1^i & t_2^i & \cdots & t_{m-1}^i & t_m^i \\
\hline
 & w_1^e & 0 & 1 & \cdots & 0 & 1 & 0 & 0 & \cdots & & \\
\textit{English Lexicon} & w_2^e & 1 & 1 & \cdots & 1 & 0 & 0 & \ddots & & & \\
 & \vdots & \multicolumn{5}{c|}{\cdots\cdots\cdots\cdots\cdots\cdots} & \vdots & & 0 & & \vdots \\
 & w_{p-1}^e & 0 & 1 & \cdots & 0 & 0 & & & & \ddots & 0 \\
 & w_p^e & 0 & 1 & \cdots & 0 & 0 & & & \cdots & 0 & 0 \\
\hline
\textit{shared names and words} & w_1^{e/i} & \mathbf{1} & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \mathbf{1} \\
 & \vdots & \multicolumn{5}{c|}{\cdots\cdots\cdots\cdots\cdots\cdots} & \multicolumn{5}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \\
\hline
\textit{common etymology} & w_1^{etym} & 0 & \mathbf{1} & \cdots & 0 & 0 & 0 & 0 & \cdots & \mathbf{1} & 0 \\
 & \vdots & \multicolumn{5}{c|}{\cdots\cdots\cdots\cdots\cdots\cdots} & \multicolumn{5}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \\
\hline
 & w_1^i & 0 & 0 & \cdots & & & 0 & 1 & \cdots & 1 & 1 \\
\textit{Italian Lexicon} & w_2^i & 0 & \ddots & & & & 1 & 1 & \cdots & 0 & 1 \\
 & \vdots & \vdots & & 0 & & \vdots & \multicolumn{5}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \\
 & w_{q-1}^i & & & & \ddots & 0 & 0 & 1 & \cdots & 0 & 1 \\
 & w_q^i & & & \cdots & 0 & 0 & 0 & 1 & \cdots & 1 & 0
\end{array}\right]$$

Figure 2: Multilingual word-by-document matrix

3 Cross Language Text Categorization

Text categorization (also text classification), “the task of automatically sorting a set of documents into categories (or classes or topics) from a predefined set” (Sebastiani, 2005), allows for the quick selection of documents from the same domain, or the same topic. It is a very well research area, dating back to the 60s (Borko and Bernick, 1962). The most frequently, and successfully, used document representation is the bag-of-words (BoWs). Results using this representation achieve accuracy in the 90%s. Most variations include feature filtering or weighing, and variations in learning algorithms (Sebastiani, 2005).

Within the area of cross-language text categorization (CLTC) several methods have been explored for producing the model for a target language L_t using information and data from the source language L_s. In a precursor task to CLTC, cross language information retrieval (CLIR), Dumais et al. (1997) find semantic correspondences in parallel (different language) corpora through latent semantic analysis (LSA). Most CLTC methods rely heavily on machine translation (MT). MT has been used: to cast the cross-language text categorization problem to the monolingual setting (Fortuna and Shawe-Taylor, 2005); to cast the cross-language text categorization problem into two monolingual settings for active learning (Liu et al., 2012); to translate and adapt a model built on language L_s to language L_t (Rigutini et al., 2005), (Shi et al., 2010); to produce parallel corpora for multi-view learning (Guo and Xiao, 2012). Wan et al. (2011) also use machine translation, but enhance the processing through domain adaptation by feature weighing, assuming that the training data in one language and the test data in the other come from different domains, or can exhibit different linguistic phenomena due to linguistic and cultural differences. Prettenhofer and Stein (2010) use a word translation oracle to produce *pivots* – pairs of semantically similar words – and use the data partitions induced by these words to find cross language structural correspondences.

In a computationally lighter framework, not dependent on MT, Gliozzo and Strapparava (2006) and Wu et al. (2008) use bilingual lexicons and aligned WordNet synsets to obtain shared features between the training data in language L_s and the testing data in language L_t. Gliozzo and Strapparava (2005), the first to use comparable as opposed to parallel corpora for CLTC, use LSA to build multilingual domain models.

The bag-of-word document representation maps a document d_i from a corpus D into a k-dimensional space $\mathbb{R}^k$, where k is the dimension of the (possibly filtered) vocabulary of the corpus: $W = \{w_1, ..., w_k\}$. Position j in the vector representation of d_i corresponds to word w_j, and it may have different values, among the most commonly used being: binary values – w_j appears (1) or not (0) in d_i; frequency of occurrence of w_j in d_i, absolute or normalized (relative to the size of the document or the size of the vocabulary); the $tf * idf(w_j, d_i, D)$.

For the task of cross language text categorization, the problem of sharing a model across languages is that the dimensions, a.k.a the vocabulary, of the two languages are largely different. Limited overlap can be achieved through shared names and words. As we have seen in the literature review, machine translation and bilingual dictionaries can be used to cast these dimensions from the source language L_s to the target language L_t. In this work we explore expanding the shared dimensions through word etymologies. Figure 2 shows schematically the binary k dimensional representation for English and Italian data, and shared dimensions.

Cross language text categorization could be used to obtain comparable corpora for building translation models. In such a situation, relying on a framework that itself relies on machine translation is not helpful. Bilingual lexicons are available for frequently studied languages, but less so for those poorer in resources. Considering such shortcomings, we look into additional linguistic information, in particular word etymology. This information impacts the data representation, by introducing new shared features between the different language corpora without the need for translation or other forms of mapping. The newly produced representation can be used in conjunction with any of the previously proposed algorithms.

Word etymologies are a novel source of linguistic information in NLP, possibly because resources that capture this information in a machine readable format are also novel. Fang et al. (2009) used limited etymological information extracted from the Collins English Dictionary (CED) for text categorization on the British National Corpus (BNC): information on the provenance of words (ranges of probability distribution of etymologies in different versions of Latin – New Latin, Late Latin, Medieval Latin) was used in a "home-made" range classifier.

The experiments presented in this paper use the bag-of-word document representation with absolute frequency values. To this basic representation we add word etymological ancestors and run classification experiments. We then use LSA – previously shown by (Dumais et al., 1997) and (Gliozzo and Strapparava, 2005) to be useful for this task – to induce the latent semantic dimensions of documents and words respectively, hypothesizing that word etymological ancestors will lead to semantic dimensions that transcend language boundaries. The vectors obtained through LSA (on the training data only) for words that are shared by the English training data and the Italian test data (names, and most importantly, etymological ancestors of words in the original documents) are then used for re-representing the training and test data. The same process is applied for Italian training and English test data. Classification is done using support vector machines (SVMs).

3.1 Data

The data we work with consists of comparable corpora of news articles in English and Italian. Each news article is annotated with one of the four categories: *culture_and_school*, *tourism*, *quality_of_life*, *made_in_Italy*. Table 1 shows the dataset statistics. The average document length is approximately 300 words.

3.2 Raw cross-lingual text categorization

As is commonly done in text categorization (Sebastiani, 2005), the documents in our data are represented as bag-of-words, and classification is done using support vector machines (SVMs).

One experimental run consists of 4 binary experiments – one class versus the rest, for each of the 4 classes. The results are reported through micro-averaged precision, recall and F1-score for the targeted class, as well as overall accuracy. The high results, on a par with text categorization experiments in the field, validates our experimental set-up.

For the cross language categorization experiments described in this paper, we use the data described above, and train on one language (English/Italian), and test on the other, using the same

	English			Italian		
Categories	Training	Test	Total	Training	Test	Total
quality_of_life	5759	1989	7748	5781	1901	7682
made_in_Italy	5711	1864	7575	6111	2068	8179
tourism	5731	1857	7588	6090	2015	8105
culture_and_school	3665	1245	4910	6284	2104	8388
Total	20866	6955	27821	24266	8088	32354

Table 1: Dataset statistics

monolingual BoW categorization

	Prec	Rec	F1	Acc
Train EN / Test EN	0.92	0.92	0.92	0.96
Train IT / Test IT	0.94	0.94	0.94	0.97

Table 2: Performance for monolingual raw text categorization

experimental set-up as for the monolingual scenario (4 binary problems). The categorization baseline (BoW_baseline in Figure 4) was obtained in this set-up. This baseline is higher than the random baseline or the positive class baseline[2] (all instances are assigned the target class in each of the 4 binary classification experiments) due to shared words and names between the two languages.

3.3 Enriching the bag-of-word representation with word etymology

As personal experience has shown us that etymological information is useful for comprehending a text in a different language, we set out to test whether this information can be useful in an automatic processing setting. We first verified whether the vocabularies of our two corpora, English and Italian, have shared word etymologies. Relying on word etymologies from the Etymological dictionary, we found that from our data's vocabulary, 518 English terms and 543 Italian terms shared 490 direct etymological ancestors. Etymological ancestors also help cluster related terms within one language – 887 etymological ancestors for 4727 English and 864 ancestors for 5167 Italian terms. This overlap further increases when adding derived forms (through the *has_derived_form* relation). The fact that this overlap exists strengthens the motivation to try using etymological ancestors for the task of text categorization.

In this first step of integrating word etymology into the experiment, we extract for each word in each document in the dataset its ancestors from the Etymological dictionary. Because each word w_j in a document d_i has associated an absolute frequency value f_{ij} (the number of occurrences of w_j in d_i), for the added etymological ancestors e_k in document D_i we associate as value the sum of frequencies of their etymological children in d_i:

$$f_{ie_k} = \sum_{\substack{w_j \in d_i \\ w_j \text{etymology } e_k}} f_{ij}$$

We make the depth of extraction a parameter, and generate data representation when considering only direct etymological antecedents (depth 1) and then up to a distance of N. For our dataset we noticed that the representation does not change after N=4, so this is the maximum depth we consider. The bag-of-words representation for each document is expanded with the corresponding etymological features.

expansion	training data vocabulary size	vocabulary overlap with testing
Train EN /Test IT		
raw	71122	14207 (19.9%)
depth 1	78936	18275 (23.1%)
depth 2	79068	18359 (23.2%)
depth 3	79100	18380 (23.2%)
depth 4	79103	18382 (23.2%)
Train IT /Test EN		
raw	78750	14110 (17.9%)
depth 1	83656	18682 (22.3%)
depth 2	83746	18785 (22.4%)
depth 3	83769	18812 (22.5%)
depth 4	83771	18814 (22.5%)

Table 3: Feature expansion with word etymologies

Table 3 shows the training data vocabulary size and increase in the overlap between the training and test data with the addition of etymological fea-

[2] In this situation the random and positive class baseline are the same: 25% F1 score.

tures. The increase is largest when introducing the immediate etymological ancestors, of approximately 4000 new (overlapping) features for both combinations of training and testing. Without etymological features the overlap was approximately 14000 for both configurations. The results obtained with this enriched BoW representation for etymological ancestor depth 1, 2 and 3 are presented in Figure 4.

3.4 Cross-lingual text categorization in a latent semantic space adding etymology

Shared word etymologies can serve as a bridge between two languages as we have seen in the previous configuration. When using shared word etymologies in the bag-of-words representation, we only take advantage of the shallow association between these new features and the classes within which they appear. But through the co-occurrence of the etymological features and other words in different documents in the training data, we can induce a deeper representation for the words in a document, that captures better the relationship between the features (words) and the classes to which the documents belong. We use latent semantic analysis (LSA) (Deerwester et al., 1990) to perform this representational transformation. The process relies on the assumption that word co-occurrences across different documents are the surface manifestation of shared semantic dimensions. Mathematically, the $\langle$word $\times$ document$\rangle$ matrix D is expressed as a product of three matrices:

$$D = V\Sigma U^T$$

by performing singular value decomposition (SVD). V would correspond roughly to a $\langle$word $\times$ latent semantic dimension$\rangle$ matrix, U^T is the transposed of a $\langle$document $\times$ latent semantic dimension$\rangle$ matrix, and Σ is a diagonal matrix whose values are indicative of the "strength" of the semantic dimensions. By reducing the size of Σ, for example by selecting the dimensions with the top K values, we can obtain an approximation of the original matrix $D \approx D_K = V_K \Sigma_K U_K^T$, where we restrict the latent semantic dimensions taken into account to the K chosen ones. Figure 3 shows schematically the process.

We perform this decomposition and dimension reduction step on the $\langle$word $\times$ document$\rangle$ matrix built from the training data only, and using K=400. Both the training and test data are then re-represented through the new word vectors from matrix V_K. Because the LSA space was built only from the training data, only the shared words and shared etymological ancestors are used to produce representations of the test data. The categorization is done again with SVM. The results of this experiment are shown in Figure 4, together with an LSA baseline – using the raw data and relying on shared words and names as overlap.

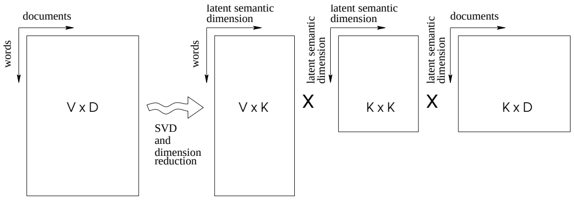
documents
words
V x D
SVD
and
dimension
reduction
latent semantic dimension
words
V x K
X
latent semantic dimension
latent semantic dimension
K x K
X
latent semantic dimension
documents
K x D

Figure 3: Schematic view of LSA

4 Discussion

The experiments whose results we present here were produced using unfiltered data – all words in the datasets, all etymological ancestors up to the desired depth, no filtering based on frequency of occurrence. Feature filtering is commonly done in machine learning when the data has many features, and in text categorization when using the bag-of-words representation in particular. We chose not to perform this step for two main reasons: (i) filtering is sensitive to the chosen threshold; (ii) LSA thrives on word co-occurrences, which would be drastically reduced by word removal. The point that etymology information is a useful addition to the task of cross-language text categorization can be made without finding the optimal filtering set-up.

The baseline experiments show that despite the relatively large word overlap (approx. 14000 terms), cross-language text categorization gives low results. Adding a first batch of etymological information – approximately 4000 shared immediate ancestors – leads to an increase of 18 points in terms of F1-score on the BoW experimental set-up for English training/Italian testing, and 21 points for Italian training/English testing. Further additions of etymological ancestors at depths 2 and 3 results in an increase of 21 points in terms of F1-score for English training/Italian testing, and 27 points for Italian training/English testing. The higher increase in performance on this experimental configuration for Italian training/English testing is explained by the higher term overlap be-

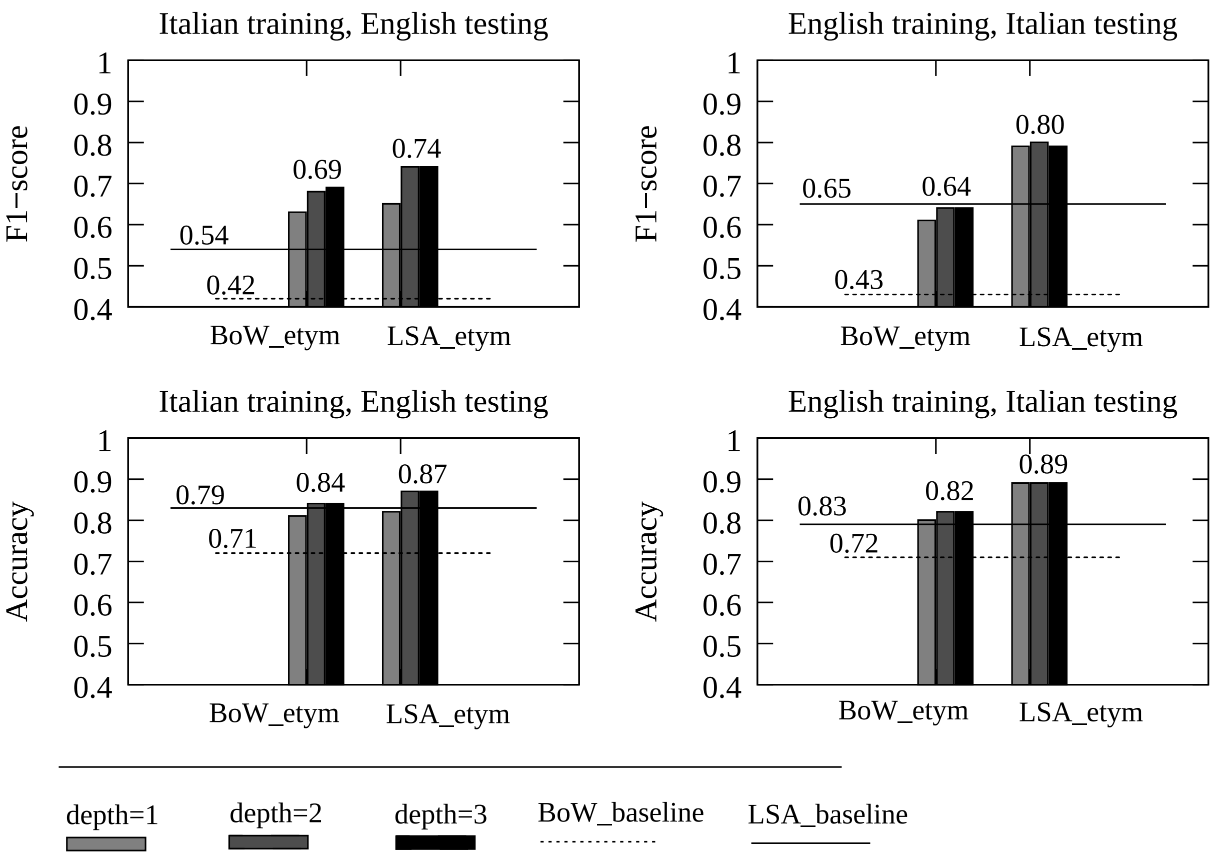

Figure 4: CLTC results with etymological features

tween the training and test data, as evidenced by the statistics in Table 3.

The next processing step induced a representation of the shared words that encodes deeper level dependencies between words and documents based on word co-occurrences in documents. The LSA space built on the training data leads to a vector representation of the shared words, including the shared etymological ancestors, that captures more than the obvious word-document co-occurrences. Using this representation leads to a further increase of 15 points in F1-score for English training/Italian testing set-up over the BoW representation, and 14 points over the baseline LSA-based categorization. The increase for the Italian training/English testing is 5 points over the BoW representation, but 20 points over the baseline LSA. We saw that the high performance BoW on Italian training/English testing is due to the high term overlap. The clue to why the increase when using LSA is lower than for English training/Italian testing is in the way LSA operates – it relies heavily on word co-occurrences in finding the latent semantic dimensions of documents and words. We expect then that in the Italian training collection, words are "less shared" among documents, which means a lower average document frequency. Figure 5 shows the changes in average document frequency for the two training collections, starting with the raw data (depth 0), and with additional etymological features.

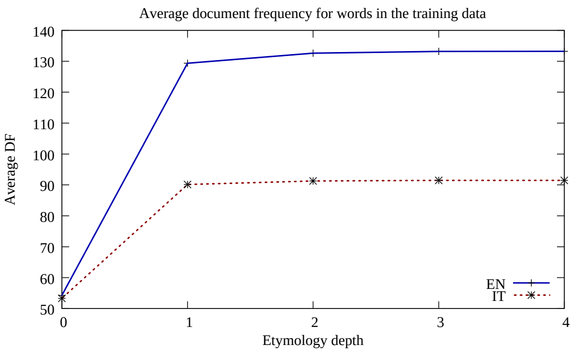

Figure 5: Document frequency changes with the addition of etymological features

The shape of the document frequency curves mirror the LSA results – the largest increase is the effect of adding the set of direct etymological ancestors, and additions of further, more distant, ancestors lead to smaller improvements.

We have performed the experiments described above on two releases of the Etymological dictionary. The results described in the paper were obtained on the latest release (February 2013). The difference in results on the two dictionary versions was significant: a 4 and 5 points increase respectively in micro-averaged F1-score in the bag-of-words setting for English training/Italian testing and Italian training/English testing, and a 2 and 6 points increase in the LSA setting. This indicates that more etymological information is better, and the dynamic nature of Wikipedia and the Wiktionary could lead to an ever increasing and better etymological resource for NLP applications.

5 Conclusion

The motivation for this work was to test the hypothesis that information about word etymology is useful for computational approaches to language, in particular for text classification. Cross-language text classification can be used to build comparable corpora in different languages, using a single language starting point, preferably one with more resources, that can thus spill over to other languages. The experiments presented have shown clearly that etymological ancestors can be used to provide the necessary bridge between the languages we considered – English and Italian. Models produced on English data when using etymological information perform with high accuracy (89%) and high F1-score (80) on Italian test data, with an increase of almost 40 points over a simple bag-of-words model, which, for crossing language boundaries, relies exclusively on shared names and words. Training on Italian data and testing on English data performed almost as well (87% accuracy, 75 F1-score). We plan to expand our experiments to more languages with shared etymologies, and investigate what characteristics of languages and data indicate that etymological information is beneficial for the task at hand.

We also plan to explore further uses for this language bridge, at a finer semantic level. Monolingual and cross-lingual textual entailment in particular would be interesting applications, because they require finding shared meaning on two text fragments. Word etymologies would allow recognizing words with shared ancestors, and thus with shared meaning, both within and across languages.

Acknowledgements

We thank the reviewers for the helpful comments. This work was financially supported by the EC-funded project EXCITEMENT – EXploring Customer Interactions through Textual EntailMENT FP7 ICT-287923. Carlo Strapparava was partially supported by the PerTe project (Trento RISE).

References

Harold Borko and Myrna Bernick. 1962. *Automatic Document Classification*. System Development Corporation, Santa Monica, CA.

Gerard de Melo and Gerhard Weikum. 2010. Towards universal multilingual knowledge bases. In *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)*, pages 149–156, New Delhi, India.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Symposium on CrossLanguage Text and Speech Retrieval*.

Alex Chengyu Fang, Wanyin Li, and Nancy Ide. 2009. Latin etymologies as features on BNC text categorization. In *23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*, pages 662–669.

Blaz Fortuna and John Shawe-Taylor. 2005. The use of machine translation tools for cross-lingual text mining. In *Learning with multiple views – Workshop at the 22nd International Conference on Machine Learning (ICML 2005)*.

Alfio Gliozzo and Carlo Strapparava. 2005. Cross language text categorization by acquiring multilingual domain models from comparable corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 553–560, Sydney, Australia.

Yuhong Guo and Min Xiao. 2012. Cross language text classification via subspace co-regularized multi-view learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland, UK.

Henry Hitchings. 2009. *The Secret Life of Words: How English Became English*. John Murray Publishers.

Yue Liu, Lin Dai, Weitao Zhou, and Heyan Huang. 2012. Active learning for cross language text categorization. In *Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2012)*, pages 195–206, Kuala Lumpur, Malaysia.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1118–1127, Uppsala, Sweden.

Leonardo Rigutini, Marco Maggini, and Bing Liu. 2005. An EM based training algorithm for cross-language text categorization. In *Proceedings of the International Conference on Web Intelligence (WI 2005)*, pages 200–206, Compiegne, France.

Fabrizio Sebastiani. 2005. Text categorization. In Alessandro Zanasi, editor, *Text Mining and its Applications*, pages 109–129. WIT Press, Southampton, UK.

Lei Shi, Rada Mihalcea, and Minhgjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1057–1067, Uppsala, Sweden.

Chang Wan, Rong Pan, and Jifei Li. 2011. Bi-weighting domain adaptation for cross-language text classification. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 1535–1540, Barcelona, Catalonia, Spain.

Ke Wu, Xiaolin Wang, and Bao-Liang Lu. 2008. Cross language text categorization using a bilingual lexicon. In *Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 165–172, Hyderabad, India.