# Metaphor Detection through Term Relevance

**Marc Schulder**
Spoken Language Systems
Saarland University
Saarbrücken, Germany
marc.schulder@lsv.uni-saarland.de

**Eduard Hovy**
Language Technologies Institute
Carnegie Mellon University
Pittsburth, PA, USA
hovy@cs.cmu.edu

## Abstract

Most computational approaches to metaphor detection try to leverage either conceptual metaphor mappings or selectional preferences. Both require extensive knowledge of the mappings/preferences in question, as well as sufficient data for all involved conceptual domains. Creating these resources is expensive and often limits the scope of these systems.

We propose a statistical approach to metaphor detection that utilizes the rarity of novel metaphors, marking words that do not match a text's typical vocabulary as metaphor candidates. No knowledge of semantic concepts or the metaphor's source domain is required.

We analyze the performance of this approach as a stand-alone classifier and as a feature in a machine learning model, reporting improvements in $F_1$ measure over a random baseline of 58% and 68%, respectively. We also observe that, as a feature, it appears to be particularly useful when data is sparse, while its effect diminishes as the amount of training data increases.

## 1 Introduction

Metaphors are used to replace complicated or unfamiliar ideas with familiar, yet unrelated concepts that share an important attribute with the intended idea. In NLP, detecting metaphors and other non-literal figures of speech is necessary to interpret their meaning correctly. As metaphors are a productive part of language, listing known examples is not sufficient. Most computational approaches to metaphor detection are based either on the theory of conceptual mappings (Lakoff and Johnson, 1980) or that of preference violation (Wilks, 1978).

Lakoff and Johnson (1980) showed that metaphors have underlying mappings between two conceptual domains: The figurative *source* domain that the metaphor is taken from and the literal *target* domain of the surrounding context in which it has to be interpreted. Various metaphors can be based on the same conceptual metaphor mapping, e.g. both *"The economy is a **house of cards**"* and *"the **stakes** of our debates appear small"* match POLITICS IS A GAME.

Another attribute of metaphors is that they violate semantic *selectional preferences* (Wilks, 1978). The theory of selectional preference observes that verbs constrain their syntactic arguments by the semantic concepts they accept in these positions. Metaphors violate these constraints, combining incompatible concepts.

To make use of these theories, extensive knowledge of pairings (either mappings or preferences) and the involved conceptual domains is required. Especially in the case of conceptual mappings, this makes it very difficult for automated systems to achieve appropriate coverage of metaphors. Even when limited to a single target domain, detecting all metaphors would require knowledge of many metaphoric source domains to cover all relevant mappings (which themselves have to be known, too). As a result of this, many systems attempt to achieve high precision for specific mappings, rather than provide general coverage.

Many approaches (Gedigian et al., 2006; Krishnakumaran and Zhu, 2007; Mohler et al., 2013; Tsvetkov et al., 2013, and more) make use of manually crafted knowledge bases such as WordNet or FrameNet to establish concept domains. Other recent works establish domains via topic modeling (Shutova et al., 2010; Heintz et al., 2013), ad-hoc clustering (Strzalkowski et al., 2013) or by using semantic similarity vectors (Hovy et al., 2013).

We introduce term relevance as a measure for how *"out of place"* a word is in a given con-



*Proceedings of the Second Workshop on Metaphor in NLP*, pages 18–26,
Baltimore, MD, USA, 26 June 2014. 

text. Our hypothesis is that words will often be out of place because they are not meant literally, but rather metaphorically. Term relevance is based on term frequency measures for target domains and mixed-domain data. The advantage of this approach is that it only requires knowledge of a text's literal target domain, but none about any source domains or conceptual mappings. As it does not require sentence structure information, it is also resistant to noisy data, allowing the use of large, uncurated corpora. While some works that utilize domain-mappings circumvent the need for pre-existing source data by generating it themselves (Strzalkowski et al., 2013; Mohler et al., 2013), our approach is truly source-independent.

We present a threshold classifier that uses term relevance as its only metric for metaphor detection. In addition we evaluate the impact of term relevance at different training sizes.

Our contributions are:

- We present a measure for non-literalness that only requires data for the literal domain(s) of a text.
- Our approach detects metaphors independently of their source domain.
- We report improvements for $F_1$ of 58% (stand-alone) and 68% (multi-feature) over a random baseline.

## 2 Term Relevance

We hypothesize that novel metaphoric language is marked by its unusualness in a given context. There will be a clash of domains, so the vocabulary will be noticeably different[1]. Therefore, an unusual choice of words may indicate metaphoricity (or non-literalness, at the least).

We measure this fact through a domain-specific *term relevance* metric. The metric consists of two features: *Domain relevance*, which measures whether a term is typical for the literal target domain of the text, and *common relevance*, which indicates terms that are so commonly used across domains that they have no discriminative power. If a term is not typical for a text's domain (i.e. has a low relevance), but is not very common either, it is considered a metaphor candidate. This can of course be extended to multiple literal domains (e.g. a political speech on fishing regulations will have both governance and maritime vocabulary), in which case a word is only considered as a metaphor if it is untypical for all domains involved.

### 2.1 Metric

We base *domain relevance* on TF-IDF (*term frequency inverse document frequency*), which is commonly used to measure the impact of a term on a particular document. Terms with a great impact receive high scores, while low scores are assigned to words that are either not frequent in the document or otherwise too frequent among other documents.

We adapt this method for *domain relevance (dr)* by treating all texts of a domain as a single "document". This new *term frequency inverse* ***domain*** *frequency* measures the impact of a term on the domain.

$$\mathrm{tf}_{dom}(t,d) = \frac{\text{\# of term } t \text{ in domain } d}{\text{\# of terms in domain } d} \quad (1)$$

$$\mathrm{idf}_{\mathrm{dom}}(t) = \log \frac{\text{\# of domains}}{\text{\# of domains containing } t} \quad (2)$$

$$\mathrm{dr}(t,d) = \mathrm{tf}_{\mathrm{dom}}(t,d) \times \mathrm{idf}_{\mathrm{dom}}(t) \quad (3)$$

To detect metaphors, we look for terms with low scores in this feature. However, due to the nature of TF-IDF, a low score might also indicate a word that is common among all domains. To filter out such candidates, we use normalized *document frequency* as a *common relevance* indicator.

$$\mathrm{cr}(t) = \frac{\text{\# of documents containing } t}{\text{\# of documents}} \quad (4)$$

In theory, we could also use *domain frequency* to determine common relevance, as we already compute it for domain relevance. However, as this reduces the feature's granularity and otherwise behaves the same (as long as domains are of equal size), we keep regular document frequency.

### 2.2 Generating Domains

We need an adequate number of documents for each domain of interest to compute *domain relevance* for it. We require specific data for the literal domain(s) of a text, but none for the metaphor's

[1] Strongly conventionalized metaphors will not meet this expectation, as they have become part of the target domain's vocabulary. Such metaphors can be easily detected by conventional means, such as knowledge bases. Our concern is therefore focused on novel metaphors.

source domains. This reduces the required number of domain data sets significantly without ruling out any particular metaphor mappings.

We extract domain-specific document collections from a larger general corpus, using the keyword query search of Apache Lucene[2], a software for indexed databases. The keywords of the query search are a set of seed terms that are considered typical literal terms for a domain. They can be manually chosen or extracted from sample data. For each domain we extract the 10,000 highest ranking documents and use them as the domain's dataset.

Afterwards, all remaining documents are randomly assigned to equally sized pseudo-domain datasets. These pseudo-domains allow us to compute the inverse of the *domain frequency* for the TF-IDF without the effort of assigning all documents to proper domains. The *document frequency* score that will be used as *common relevance* is directly computed on the documents of the complete corpus.

## 3 Data

We make use of two different corpora. The first is the domain-independent corpus required for computing term relevance. The second is an evaluation corpus for the *governance* domain on which we train and test our systems.

Both corpora are preprocessed using NLTK (Loper and Bird, 2002)[3]. After tokenization, stopwords and punctuation are removed, contractions expanded (e.g. *we've* to *we have*) and numbers generalized (e.g. *1990's* to *@'s*). The remaining words are reduced to their stem to avoid data sparsity due to morphological variation.

In case of the domain corpus, we also removed generic web document contents, such as HTML mark-up, JavaScript/CSS code blocks and similar boilerplate code[4].

### 3.1 Domain Corpus

As a basis for term relevance, we require a large corpus that is domain-independent and ideally also style-independent (i.e. not a newspaper corpus or Wikipedia). The world wide web meets these requirements. However, we cannot use public online search engines, such as Google or Bing, because they do not allow a complete overview of their indexed documents. As we require this provide to generate pseudo-domains and compute the inverse document/domain frequencies, we use a precompiled web corpus instead.

*ClueWeb09*[5] contains one billion web pages, half of which are English. For reasons of processing time and data storage, we limited our experiments to a single segment (`en0000`), containing 3 million documents. The time and storage considerations apply to the generation of term relevance values during preprocessing, due to the requirements of database indexing. They do not affect the actual metaphor detection process, therefore, we do not expect scalability to be an issue. As ClueWeb09 is an unfiltered web corpus, spam filtering was required. We removed 1.2 million spam documents using the *Waterloo Spam Ranking for ClueWeb09*[6] by Cormack et al. (2011).

### 3.2 Evaluation Corpus

Evaluation of the two classifiers is done with a corpus of documents related to the concept of *governance*. Texts were annotated for metaphoric phrases and phrases that are decidedly *in-domain*, as well as other factors (e.g. affect) that we will not concern ourselves with. The focus of annotation was to exhaustively mark metaphors, irrespective of their novelty, but avoid idioms and metonymy.

The corpus is created as part of the *MICS: Metaphor Interpretation in terms of Culturally-relevant Schemas* project by the U.S. Intelligence Advanced Research Projects Activity (IARPA). We use a snapshot containing 2,510 English sentences, taken from 312 documents. Of the 2,078 sentences that contain metaphors, 72% contain only a single metaphoric phrase. The corpus consists of around 48k tokens, 12% of which are parts of metaphors. Removing stopwords and punctuation reduces it to 23k tokens and slightly skews the distribution, resulting in 15% being metaphors.

We divide the evaluation data into 80% development and 20% test data. All reported results are based on test data. Where training data is required for model training (see section 5), ten-fold cross validation is performed on the development set.

[2] `http://lucene.apache.org/core/`

[3] `http://nltk.org`

[4] Mark-up and boilerplate removal scripts adapted from `http://love-python.blogspot.com/2011/04/html-to-text-in-python.html` and `http://effbot.org/zone/re-sub.htm`

[5] `http://lemurproject.org/clueweb09/`

[6] `http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/`

| Subdomain | Seed Terms |
|---|---|
| Executive | administer rule govern lead |
| Legislative | pass law regulate debate parliament |
| Judicial | judge hearing case rule case sentence |
| Administr. | administer manage issue permits analyze study facilitate obstruct |
| Enforcement | enforce allow permit require war make mandate defeat overcome |
| Economy | budget tax spend plan finances |
| Election | vote campaign canvass elect defeat form party create platform |
| Acceptance | government distrust (de)legitimize authority reject oppose strike flag protest pride salute march accept |

Table 1: Manually selected seed terms for document search queries. The 10k documents with the highest relevance to the seeds are assigned to the subdomain cluster.

## 4 Basic Classification

To gain an impression of the differentiating power of *tf-idf* in metaphor detection, we use a basic threshold classifier ($tc$) that uses domain relevance ($dr$) and common relevance ($cr$) as its only features. Given a word $w$, a target domain $d$ and two thresholds $\delta$ and $\gamma$:

$$tc(w,d) = \begin{cases} \textbf{metaphor} & \text{if } dr(w,d) < \delta \\ & \text{and } cr(w) < \gamma \\ \textbf{literal} & \text{otherwise} \end{cases} \quad (5)$$

In cases where a text has more than one literal domain or multiple relevant subdomains are available, a word is only declared a metaphor if it is not considered literal for any of the (sub)domains.

### 4.1 Seed Terms

The threshold classifier is evaluated using two different sets of seed terms. The first set is composed of 60 manually chosen terms[7] from eight *governance* subdomains. These are shown in table 1. Each subdomain corpus consists of its 10,000 highest ranking documents. We do not subdivide the evaluation corpus into these subdomains. Rather, we assume that each sentence belongs to all eight subdomains[8], so a word is only considered a metaphor if it is non-literal for all of them. Preliminary experiments showed that this provides better performance than using a single domain corpus with more documents.

| | | | | |
|---|---|---|---|---|
| principl | financi | legisl | congress | crisi |
| corpor | famili | middl | compani | futur |
| countri | global | negoti | medicaid | unit |
| industri | promis | polici | constitut | save |
| obama | health | creat | capitalist | hous |
| clinton | nation | dream | american | busi |
| nuclear | amend | great | medicar | care |
| econom | million | feder | recoveri | job |
| commun | potenti | polit | freedom | law |
| prosper | energi | elect | program | new |

Table 2: The fifty stems with the highest tf-idf score in the gold data. Used as seed terms for document search, generating a single *governance* domain. Stems are listed in no particular order.

As the first set of seeds is chosen without statistical basis, the resulting clusters might miss important aspects of the domain. To ensure that our evaluation is not influenced by this, we also introduce a second seed set, which is directly based on the development data. As we mentioned in section 3.2, sentences in the MICS corpus were not only annotated for metaphoric phrases, but also for such that are decidedly domain-relevant. For example in the sentence *"Our **economy** is the **strongest** on earth"*, *economy* is annotated as in-domain and *strongest* as metaphor.

Based on these annotations, we divide the entire development data into three bags of words, one each for metaphor, in-domain and unmarked words. We then compute TF-IDF values for these bags, as we did for the domain clusters. The fifty terms[9] that score highest for the in-domain bag (i.e. those that make the texts identifiable as *governance* texts) are used as the second set of seeds (table 2). It should be noted that while the seeds were based on the evaluation corpus, the resulting term relevance features were nevertheless computed using clusters extracted from the web corpus.

[7] Terms were chosen according to human understanding of typical terms for *governance*. No optimization of the term choices was performed thereafter.

[8] As our evaluation corpus does not specify secondary domains for its texts (e.g. *fishery*), we chose not to define any further domains at this point.

[9] Various sizes were tried for the seed set. Using fifty terms offered the best performance, being neither too specific nor watering down the cluster quality. It is also close to the size of our first seed set.

| | $F_1$ | Prec | Rec |
|---|---|---|---|
| **Random** | 0.222 | 0.142 | 0.500 |
| **All Metaphor** | 0.249 | 0.142 | **1.000** |
| **T-hold: Manual Seeds** | **0.350** | **0.276** | 0.478 |
| **T-hold: 50-best Seeds** | 0.346 | 0.245 | ***0.591*** |

Table 3: Summary of best performing settings for each threshold classifier model. Bold numbers indicate best performance; slanted bold numbers: best threshold classifier recall. All results are significantly different from the baselines with $p < 0.01$.

### 4.2 Evaluation

We evaluate and optimize our systems for the $F_1$ metric. In addition we provide precision and recall. Accuracy on the other hand proved an inappropriate metric, as the prevalence of literal words in our data resulted in a heavy bias. We evaluate on a token-basis, as half of the metaphoric phrases consist of a single word and less than 15% are more than three words long (including stopwords, which are filtered out later). Additionally, evaluating on a phrase-basis would have required grouping non-metaphor sections into phrases of a similar format.

Based on dev set performance, we choose a domain relevance threshold $\delta = 0.02$ and a common relevance threshold $\gamma = 0.1$. We provide a random baseline, as well as one that labels all words as metaphors, as they are the most frequently encountered baselines in related works. Results are shown in table 3.

Both seed sets achieve similar F-scores, beating the baselines by between 39% and 58%, but their precision and recall performance differs notably. Both models are significantly better than the baseline and significantly different from one another with $p < 0.01$. Significance was computed for a two-tailed $t$-test using `sigf` (Padó, 2006)[10].

Using manually chosen seed terms results in a recall rate that is slightly worse than chance, but it is made up by the highest precision. The fact that this was achieved without expert knowledge or term optimization is encouraging.

The classifier using the fifty best *governance* terms shows a stronger recall, most likely because the seeds are directly based on the development data, resulting in a domain cluster that more closely resembles the evaluation corpus. Precision, on the other hand, is slightly below that of the manual seed classifier. This might be an effect of the coarser granularity that a single domain score offers, as opposed to eight subdomain scores.

## 5 Multi-Feature Classification

Using term relevance as the only factor for metaphor detection is probably insufficient. Rather, we anticipate to use it either as a prefiltering step or as a feature for a more complex metaphor detection system. To simulate the latter, we use an off-the-shelf machine learning classifier with which we test how *term relevance* interacts with other typical word features, such as *part of speech*. As we classify all words of a sentence, we treat the task as a binary sequential labeling task.

Preliminary tests were performed with HMM, CRF and SVM classifiers. CRF performance was the most promising. We use CRFsuite (Okazaki, 2007)[11], an implementation of conditional random fields that supports continuous values via scaling factors. Training is performed on the development set using ten-fold cross validation.

We present results for bigram models. Larger n-grams were inspected, too, including models with look-ahead functionality. While they were slightly more robust with regard to parameter changes, there was no improvement over the best bigram model. Also, as metaphor processing still is a low resource task for which sufficient training data is hard to come by, bigrams are the most accessible and representative option.

### 5.1 Training Features

We experimented with different representations for the term relevance features. As they are continuous values, they could be used as continuous features. Alternatively, they could be represented as binary features, using a cut-off value as for our threshold classifier. In the end, we chose a hybrid approach where thresholds are used to create binary features, but are also scaled according to their score. Thresholds were again determined on the dev set and set to $\delta = 0.02$ and $\gamma = 0.79$.

Each domain receives an individual domain relevance feature. There is only a single common rel-

[10] `http://www.nlpado.de/~sebastian/software/sigf.shtml`

[11] `http://www.chokkan.org/software/crfsuite/`

| | $F_1$ | Prec | Rec |
|---|---|---|---|
| **All Metaphor** | 0.249 | 0.142 | **1.000** |
| **T-hold: Manual Seeds** | 0.350 | 0.276 | 0.478 |
| **CRF: Basic** | 0.187 | 0.706 | 0.108 |
| **CRF: Rel** | 0.219 | 0.683 | 0.130 |
| **CRF: PosLex** | 0.340 | **0.654** | 0.230 |
| **CRF: PosLexRel** | **0.373** | 0.640 | ***0.263*** |

Table 4: Summary of best performing settings for each CRF model. Bold numbers indicate best performance; slanted bold numbers: best CRF recall. All results are significantly different from the baseline with $p < 0.01$.

evance feature, as it is domain-independent. Surprisingly, we found no noteworthy difference in performance between the two seed sets (manual and 50-best). Therefore we only report results for the manual seeds.

In addition to term relevance, we also provide part of speech (pos) and lexicographer sense (lex) as generic features. The part of speech is automatically generated using NLTK's *Maximum Entropy POS Tagger*, which was trained on the *Penn Treebank*. To have a semantic feature to compare our relevance weights to, we include WordNet's lexicographer senses (Fellbaum, 1998), which are coarse-grained semantic classes. Where a word has more than one sense, the first was chosen. If no sense exists for a word, the word is given a `sense unknown` placeholder value.

### 5.2 Performance Evaluation

Performance of the CRF system (see table 4) seems slightly disappointing at first when compared to our threshold classifier. The best-performing CRF beats the threshold classifier by only two points of F-score, despite considerably richer training input. Precision and recall performance are reversed, i.e. the CRF provides a higher precision of 0.6, but only detects one out of four metaphor words. All models provide stable results for all folds, their standard deviation (about 0.01 for $F_1$) being almost equal to that of the baseline.

All results are significantly different from the baseline as well as from each other with $p < 0.01$, except for the precision scores of the three non-basic CRF models, which are significantly different from each other with $p < 0.05$.

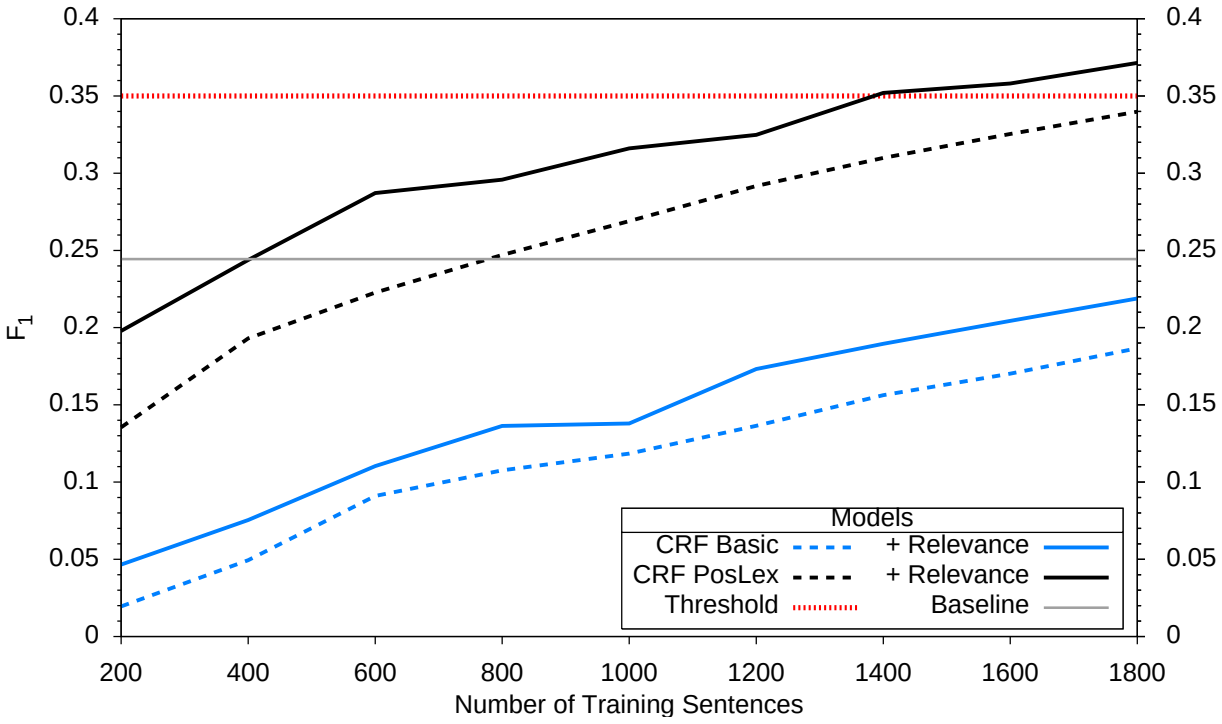


Figure 1: Performance curves for various training data sizes. Models with term relevance features (solid lines) outperform models without term relevance (dashed lines) at 1400 sentences. 1800 sentences represent the entire training set. Baseline (thin line) and best threshold classifier (dotted line) provided for reference.

Adding term relevance provides a consistent boost of 0.025 to the F-score. This boost, however, is rather marginal in comparison to the one provided by part of speech and lexicographer sense. A possible reason for this could be that the item weights learned during training correspond too closely to our term relevance scores, thus making them obsolete when enough training data is provided. The next section explores this possibility by comparing different amounts of training data.

### 5.3 Training Size Evaluation

With 2000 metaphoric sentences, the dataset we used was already among the largest annotated corpora. By reducing the amount of training data we evaluate whether term relevance is an efficient feature when data is sparse. To this end, we repeat our ten-fold cross validations, but withhold some of the folds from each training set.

Figure 1 compares the performance of CRF feature configurations with and without term relevance. In both cases adding term relevance outperforms the standard configuration's top performance with 400 sentences less, saving about a quarter of the training data.

In figure 2 we also visualize the relative gain that adding term relevance provides. As one can see, small datasets profit considerably more from our metric. Given only 200 sentences, the PosLex model receives 4.7 times the performance gain from term relevance it got at at maximum training size. The basic model has a factor of 6.8. This supports our assumption that term relevance is similar to the item weights learned during CRF training. As labeled training data is considerably more expensive to create than corpora for term relevance, this is an encouraging observation.

## 6 Related Work

For a comprehensive review on computational metaphor detection, see Shutova (2010). We limit our discussion to publications that were not covered by the review. While there are several papers evaluating on the same domain, direct comparison proved to be difficult, as many works were either evaluated on a sentence level (which our data was inappropriate for, as 80% of sentences contained metaphors) or did not provide coverage information. Another difference was that most evaluations were performed on balanced datasets, while our own data was naturally skewed for literal terms.

Strzalkowski et al. (2013) follow a related hypothesis, assuming that metaphors lack topical relatedness to in-domain words while being syntactically connected to them. Instead of using the metaphor candidate's relevance to a target domain corpus to judge relatedness, they circumvent the

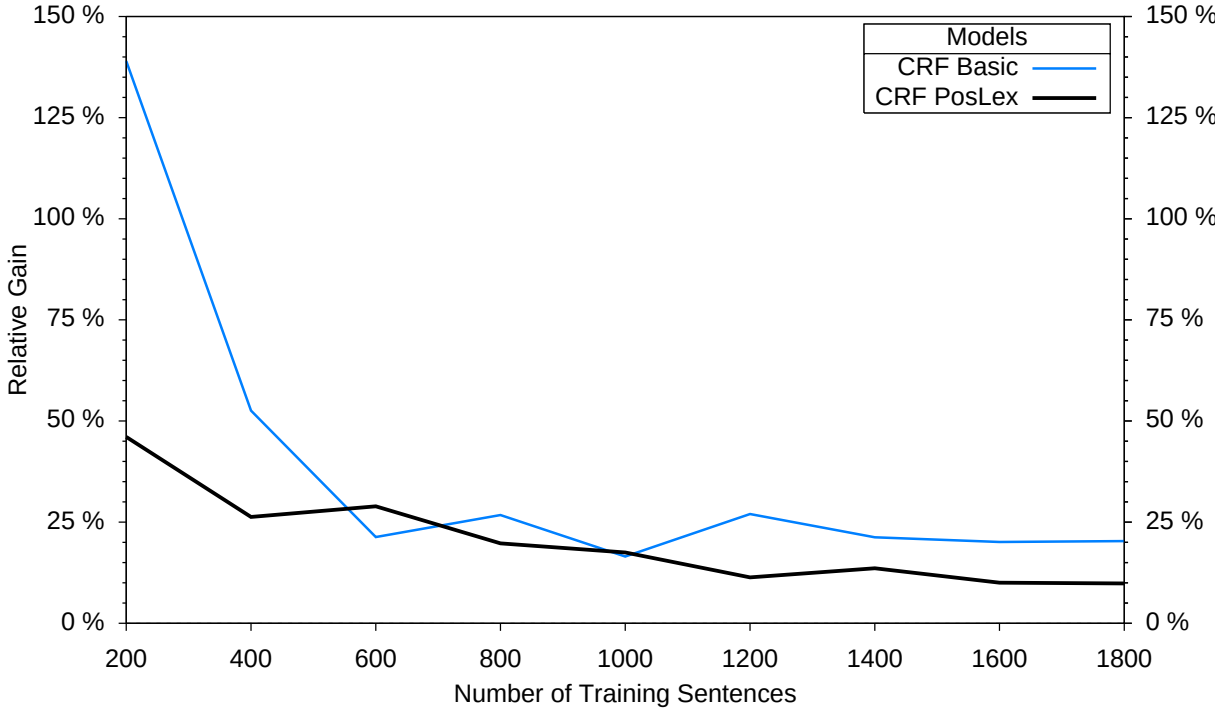


Figure 2: Relative performance gain of models obtained from addition of term relevance features.

need for pre-existing source data by generating ad-hoc collocation clusters and check whether the two highest ranked source clusters share vocabulary with the target domain. Further factors in their decision process are co-ocurrences in surrounding sentences and psycholinguistic imageability scores (i.e. how easy it is to form a mental picture of a word). Evaluating on data in the *governance* domain, they achieve an accuracy of 71% against an *all metaphor* baseline of 46%, but report no precision or recall.

Mohler et al. (2013) and Heintz et al. (2013) also evaluate on the *governance* domain. Rather than detecting metaphors at a word-level, both detect whether sentences contain metaphors. Mohler et al. (2013) compare semantic signatures of sentences to signatures of known metaphors. They, too, face a strong bias against the metaphor label and show how this can influence the balance between precision and recall. Heintz et al. (2013) classify sentences as containing metaphors if their content is related to both a target and source domain. They create clusters via topic modeling and, like us, use manually chosen seed terms to associate them with domains. Unlike our approach, theirs also requires seeds of all relevant source domains. They observe that identifying metaphors, even on a sentence level, is difficult even for experienced annotators, as evidenced by an inter-annotator agreement of $\kappa = 0.48$.

Shutova et al. (2010) use manually annotated seed sentences to generate source and target domain vocabularies via spectral clustering. The resulting domain clusters are used for selectional preference induction in verb-noun relations. They report a high precision of 0.79, but have no data on recall. Target concepts appearing in similar lexico-syntactic contexts are mapped to the same source concepts. The resulting mappings are then used to detect metaphors. This approach is notable for its combination of distributional clustering and selectional preference induction. Verbs and nouns are clustered into topics and linked through induction of selectional preferences, from which metaphoric mappings are deduced. Other works (Séaghdha, 2010; Ritter et al., 2010) use topic modeling to directly induce selectional preferences, but have not yet been applied to metaphor detection.

Hovy et al. (2013) generalize semantic preference violations from verb-noun relations to any syntactic relation and learn these in a supervised manner, using SVM and CRF models. The CRF is not the overall best-performing system, but achieves the highest precision of 0.74 against an all-metaphor baseline of 0.49. This is in line with our own observations. While they argue that metaphor detection should eventually be performed on every word, their evaluation is limited to a single expression per sentence.

Our work is also related to that of Sporleder and Li (2009) and Li and Sporleder (2010), in which they detect idioms through their lack of semantic cohesiveness with their context. Cohesiveness is measured via co-occurence of idiom candidates with other parts of a text in web searches. They do not make use of domains, basing their measure entirely on the lexical context instead.

## 7 Conclusion

We have presented term relevance as a non-literalness indicator and its use for metaphor detection. We showed that even on its own, term relevance clearly outperforms the baseline by 58% when detecting metaphors on a word basis.

We also evaluated the utility of term relevance as a feature in a larger system. Results for this were mixed, as the general performance of our system, a sequential CRF classifier, was lower than anticipated. However, tests on smaller training sets suggest that term relevance can help when data is sparse (as it often is for metaphor processing). Also, precision was considerably higher for CRF, so it might be more useful for cases where coverage is of secondary importance.

For future work we plan to reimplement the underlying idea of term relevance with different means. Domain datasets could be generated via topic modeling or other clustering means (Shutova et al., 2010; Heintz et al., 2013) and should also cover dynamically detected secondary target domains. Instead of using TF-IDF, term relevance can be modeled using semantic vector spaces (see Hovy et al. (2013)). While our preliminary tests showed better performance for CRF than for SVM, such a change in feature representation would also justify a re-evaluation of our classifier choice. To avoid false positives (and thus improve precision), we could generate ad-hoc source domains, like Strzalkowski et al. (2013) or Shutova et al. (2010) do, to detect overlooked literal connections between source and target domain.

## Acknowledgements

We would like to thank the reviewers and proofreaders for their valuable input.

This research effort was in part supported by the German Academic Exchange Service (DAAD) scholarship program PROMOS with funds from the Federal Ministry of Education and Research (BMBF), awarded by the International Office of Saarland University as well as by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0025. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DAAD, BMBF, IARPA, DoD/ARL or the German or U.S. Government.

## References

Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465.

Christiane Fellbaum. 1998. ed. WordNet: an electronic lexical database. *MIT Press, Cambridge MA*, 1:998.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the HLT/NAACL-06 Workshop on Scalable Natural Language Understanding*, pages 41–48.

Ilana Heintz, Ryan Gabbard, Mahesh Srinivasan, David Barner, Donald S Black, Marjorie Freedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphor with lda topic modeling. *Proceedings of the ACL-13 Workshop on Metaphor*, page 58.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. *Proceedings of the ACL-13 Workshop on Metaphor*, page 52.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the HLT/NAACL-07 Workshop on Computational Approaches to Figurative Language*, pages 13–20. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*, volume 111. University of Chicago Press.

Linlin Li and Caroline Sporleder. 2010. Using gaussian mixture models to detect figurative language in context. In *Proceedings of NAACL-10*, pages 297–300. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL-02 workshop on Interactive presentation sessions*, pages 63–70. Association for Computational Linguistics.

Michael Mohler, David Bracewell, David Hinote, and Marc Tomlinson. 2013. Semantic signatures for example-based linguistic metaphor detection. *Proceedings of the ACL-13 Workshop on Metaphor*, page 27.

Naoaki Okazaki, 2007. *CRFsuite: a fast implementation of conditional random fields (CRFs)*.

Sebastian Padó, 2006. *User's guide to `sigf`: Significance testing by approximate randomisation*.

Alan Ritter, Oren Etzioni, et al. 2010. A latent dirichlet allocation method for selectional preferences. In *ACT-10*, pages 424–434. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of ACL-10*, pages 435–444. Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of COLING-10*, pages 1002–1010. Association for Computational Linguistics.

Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proceedings of ACL-10*, pages 688–697. Association for Computational Linguistics.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of EACL09*, pages 754–762. Association for Computational Linguistics.

Tomek Strzalkowski, George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Boris Yamrom, Samira Shaikh, Ting Liu, Kit Cho, Umit Boz, Ignacio Cases, et al. 2013. Robust extraction of metaphors from novel data. *Proceedings of the ACL-13 Workshop on Metaphor*, page 67.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. *Proceedings of the ACL-13 Workshop on Metaphor*, page 45.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.