Automatic Idiom Identification in Wiktionary

Grace Muzny and **Luke Zettlemoyer**
Computer Science & Engineering
University of Washington
Seattle, WA 98195
{muznyg,lsz}@cs.washington.edu

Abstract

Online resources, such as Wiktionary, provide an accurate but incomplete source of idiomatic phrases. In this paper, we study the problem of automatically identifying idiomatic dictionary entries with such resources. We train an idiom classifier on a newly gathered corpus of over 60,000 Wiktionary multi-word definitions, incorporating features that model whether phrase meanings are constructed compositionally. Experiments demonstrate that the learned classifier can provide high quality idiom labels, more than doubling the number of idiomatic entries from 7,764 to 18,155 at precision levels of over 65%. These gains also translate to idiom detection in sentences, by simply using known word sense disambiguation algorithms to match phrases to their definitions. In a set of Wiktionary definition example sentences, the more complete set of idioms boosts detection recall by over 28 percentage points.

1 Introduction

Idiomatic language is common and provides unique challenges for language understanding systems. For example, a *diamond in the rough* can be the literal unpolished object or a crude but lovable person. Understanding such distinctions is important for many applications, including parsing (Sag et al., 2002) and machine translation (Shutova et al., 2012).

We use Wiktionary as a large, but incomplete, reference for idiomatic entries; individual entries can be marked as idiomatic but, in practice, most are not. Using these incomplete annotations as supervision, we train a binary Perceptron classifier for identifying idiomatic dictionary entries. We introduce new lexical and graph-based features that use WordNet and Wiktionary to compute semantic relatedness. This allows us to learn, for example, that the words in the phrase *diamond in the rough* are more closely related to the words in its literal definition than the idiomatic one. Experiments demonstrate that the classifier achieves precision of over 65% at recall over 52% and that, when used to fill in missing Wiktionary idiom labels, it more than doubles the number of idioms from 7,764 to 18,155.

These gains also translate to idiom detection in sentences, by simply using the Lesk word sense disambiguation (WSD) algorithm (1986) to match phrases to their definitions. This approach allows for scalable detection with no restrictions on the syntactic structure or context of the target phrase. In a set of Wiktionary definition example sentences, the more complete set of idioms boosts detection recall by over 28 percentage points.

2 Related Work

To the best of our knowledge, this work represents the first attempt to identify dictionary entries as idiomatic and the first to reduce idiom detection to identification via a dictionary.

Previous idiom detection systems fall in one of two paradigms: *phrase classification*, where a phrase p is always idiomatic or literal, e.g. (Gedigian et al., 2006; Shutova et al., 2010), or *token classification*, where each occurrence of a phrase p can be idiomatic or literal, e.g. (Katz and Giesbrecht, 2006;

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1417–1421,
Seattle, Washington, USA, 18-21 October 2013.

Birke and Sarkar, 2006; Li and Sporleder, 2009). Most previous idiom detection systems have focused on specific syntactic constructions. For instance, Shutova et al. (2010) consider subject/verb (*campaign surged*) and verb/direct-object idioms (*stir excitement*) while Fazly and Stevenson (2006), Cook et al. (2007), and Diab and Bhutada (2009) detect verb/noun idioms (*blow smoke*). Fothergill and Baldwin (2012) are syntactically unconstrained, but only study Japanese idioms. Although we focus on identifying idiomatic dictionary entries, one advantage of our approach is that it enables syntactically unconstrained token-level detection for any phrase in the dictionary.

3 Formal Problem Definitions

Identification For identification, we assume data of the form $\{(\langle p_i, d_i \rangle, y_i) : i = 1 \ldots n\}$ where p_i is the phrase associated with definition d_i and $y_i \in \{\text{literal, idiomatic}\}$. For example, this would include both the literal pair ⟨ "leave for dead", "To abandon a person or other living creature that is injured or otherwise incapacitated, assuming that the death of the one abandoned will soon follow."⟩ and the idiomatic pair ⟨ "leave for dead", "To disregard or bypass as unimportant." ⟩. Given $\langle p_i, d_i \rangle$, we aim to predict y_i.

Detection To evaluate identification in the context of detection, we assume data $\{(\langle p_i, e_i \rangle, y_i) : i = 1 \ldots n\}$. Here, p_i is the phrase in example sentence e_i whose idiomatic status is labeled $y_i \in \{\text{idiomatic, literal}\}$. One such idiomatic pair is ⟨"heart to heart", "They sat down and had a long overdue heart to heart about the future of their relationship."⟩. Given $\langle p_i, e_i \rangle$, we again aim to predict y_i.

4 Data

We gathered phrases, definitions, and example sentences from the English-language Wiktionary dump from November 13th, 2012.[1]

Identification Phrase, definition pairs $\langle p, d \rangle$ were gathered with the following restrictions: the title of the Wiktionary entry must be English, p must composed of two or more words w, and $\langle p, d \rangle$ must be in its base form—senses that are not defined as a different tense of a phrase—e.g. the pair ⟨ "weapons of mass destruction", "Plural form of weapon of mass destruction" ⟩ was removed while the pair ⟨ "weapon of mass destruction", "A chemical, biological, radiological, nuclear or other weapon that ... "⟩ was kept.

Data Set	Literal	Idiomatic	Total
All	56,037	7,764	63,801
Train	47,633	6,600	54,233
Unannotated Dev	2,801	388	3,189
Annotated Dev	2,212	958	3,170
Unannotated Test	5,603	776	6,379
Annotated Test	4,510	1,834	6,344

Figure 1: Number of dictionary entries with each class for the Wiktionary identification data.

Data Set	Literal	Idiomatic	Total
Dev	171	330	501
Test	360	695	1055

Figure 2: Number of sentences of each class for the Wiktionary detection data.

Each pair $\langle p, d \rangle$ was assigned label y according to the idiom labels in Wiktionary, producing the Train, Unannotated Dev, and Unannotated Test data sets. In practice, this produces a noisy assignment because a majority of the idiomatic senses are not marked. The development and test sets were annotated to correct these potential omissions. Annotators used the definition of an idiom as a "phrase with a non-compositional meaning" to produce the Annotated Dev and Annotated Test data sets. Figure 1 presents the data statistics.

We measured inter-annotator agreement on 1,000 examples. Two annotators marked each dictionary entry as literal, idiomatic, or indeterminable. Less than one half of one percent could not be determined[2]—the computed kappa was 81.85. Given this high level of agreement, the rest of the data were only labeled by a single annotator, following the methodology used with the VNC-Tokens Dataset (Cook et al., 2008).

Detection For detection, we gathered the example sentences provided, when available, for each definition used in our annotated identification data sets. These sentences provide a clean source of develop-

[1] We used the Java Wiktionary Library (Zesch et al., 2008).

[2] The indeterminable pairs were omitted from the data.

ment and test data containing idiomatic and literal phrase usages. In all, there were over 1,300 unique phrases, half of which had more than one possible dictionary definition in Wiktionary. Figure 2 provides the complete statistics.

5 Identification Model

For identification, we use a linear model that predicts class $y^* \in \{\text{literal, idiomatic}\}$ for an input pair $\langle p, d \rangle$ with phrase p and definition d. We assign the class:

$$y^* = \arg\max_y \theta \cdot \phi(p, d, y)$$

given features $\phi(p, d, y) \in \mathbb{R}^n$ with associated parameters $\theta \in \mathbb{R}^n$.

Learning In this work, we use the averaged Perceptron algorithm (Freund and Schapire, 1999) to perform learning, which was optimized in terms of iterations T, bounded by range [1, 100], by maximizing F-measure on the development set.

The models described correspond to the features they use. All models are trained on the same, unannotated training data.

Features The features that were developed fall into two categories: lexical and graph-based features. The lexical features were motivated by the intuition that literal phrases are more likely to have closely related words in d to those in p because literal phrases do not break the principle of compositionality. All words compared are stemmed versions. Let $count(w, t)$ = number of times word w appears in text t.

- synonym overlap: Let S be the set of synonyms as defined in Wiktionary for all words in p. Then, we define the synonym overlap = $\frac{1}{|S|} \sum_{s \in S} count(s, d)$.
- antonym overlap: Let A be the set of antonyms as defined in Wiktionary for all words in p. Then, we define the antonym overlap = $\frac{1}{|A|} \sum_{a \in A} count(a, d)$.
- average number of capitals:[3] The value of $\frac{\text{number of capital letters in } p}{\text{number of words in } p}$.

[3] In practice, this feature identifies most proper nouns.

Graph-based features use the graph structure of WordNet 3.0 to calculate path distances. Let $distance(w, v, rel, n)$ be the minimum distance via links of type rel in WordNet from a word w to a word v, up to a threshold max integer value n, and 0 otherwise. The features compute:

- closest synonym:

$$\min_{w \in p, v \in d} distance(w, v, synonym, 5)$$

- closest antonym:[4]

$$\min_{w \in p, v \in d} distance(w, v, antonym, 5)$$

- average synonym distance:

$$\frac{1}{|p|} \sum_{w \in p, v \in d} distance(w, v, synonym, 5)$$

- average hyponym:

$$\frac{1}{|p|} \sum_{w \in p, v \in d} distance(w, v, hyponym, 5)$$

- synsets connected by an antonym: This feature indicates whether the following is true. The set of synsets Syn_p, all synsets from all words in p, and the set of synsets Syn_d, all synsets from all words in d, are connected by a shared antonym. This feature follows an approach described by Budanitsky et al. (2006).

6 Experiments

We report identification and detection results, varying the data labeling and choice of feature sets.

6.1 Identification

Random Baseline We use a proportionally random baseline for the identification task that classifies according to the proportion of literal definitions seen in the training data.

Results Figure 3 provides the results for the baseline, the full approach, and variations with subsets of the features. Results are reported for the original, unannotated test set, and the same test examples with corrected idiom labels. All models increased

[4] The first relation expanded was the antonym relation. All subsequent expansions were via synonym relations.

Data Set	Model	Rec.	Prec.	F1
Unannotated	Lexical	85.8	21.9	34.9
	Graph	62.4	26.6	37.3
	Lexical+Graph	70.5	28.1	**40.1**
	Baseline	12.2	11.9	12.0
Annotated	Lexical	81.2	49.3	61.4
	Graph	64.3	51.3	57.1
	Lexical+Graph	75.0	52.9	**62.0**
	Baseline	29.5	12.5	17.6

Figure 3: Results for idiomatic definition identification.

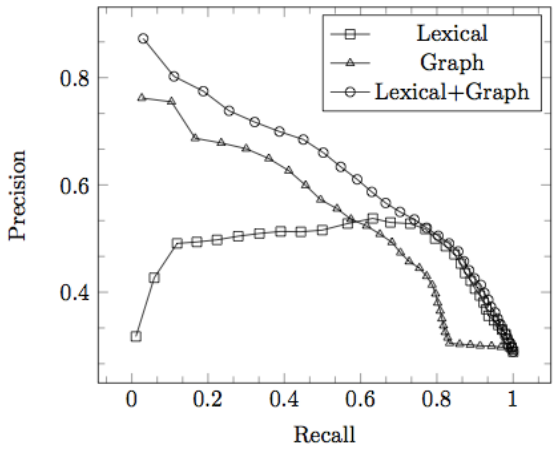

Figure 4: Precision and recall with varied features on the annotated test set.

over their corresponding baselines by more than 22 points and both feature families contributed.[5]

Figure 4 shows the complete precision, recall curve. We selected our operating point to optimize F-measure, but we see that the graph features perform well across all recall levels and that adding the lexical features provides consistent improvement in precision. However, other points are possible, especially when aiming for high precision to extend the labels in Wiktionary. For example, the original 7,764 entries can be extended to 18,155 at 65% precision, 9,594 at 80%, or 27,779 at 52.9%.

Finally, Figures 5 and 6 present qualitative results, including newly discovered idioms and high scoring false identifications. Analysis reveals where our system has room to improve—errors most often occur with phrases that are specific to a certain field, such as sports or mathematics, and with phrases whose words also appear in their definitions.

[5]We also ran ablations demonstrating that removing each feature from the Lexical+Graph model hurt performance, but omit the detailed results for space.

Phrase	Definition
feel free	You have my permission.
live down	To get used to something shameful.
nail down	To make something (e.g. a decision or plan) firm or certain.
make after	To chase.
get out	To say something with difficulty.
good riddance to bad rubbish	A welcome departure.
as all hell	To a great extent or degree; very.
roll around	To happen, occur, take place.

Figure 5: Newly discovered idioms.

Phrase	Definition
put asunder	To sunder; disjoin; separate; disunite; divorce; annul; dissolve.
add up	To take a sum.
peel off	To remove (an outer layer or covering, such as clothing).
straighten up	To become straight, or straighter.
wild potato	The edible root of this plant.
shallow embedding	The act of representing one logic or language with another by providing a syntactic translation.

Figure 6: High scoring false identifications.

6.2 Detection

Approach We use the Lesk (1986) algorithm to perform WSD, matching an input phrase p from sentence e to the definition d in Wiktionary that defines the sense p is being used in. The final classification y is then assigned to $\langle p, d \rangle$ by the identification model.

Results Figure 7 shows detection results. The baseline for this experiment is a model that assigns the default labels within Wiktionary to the disambiguated definition. The Annotated model is the Lexical+Graph model shown in Figure 3 evaluated on the annotated data. The +Default setting augments the identification model by labeling the $\langle p, e \rangle$ as idiomatic if either the model or the original label within Wiktionary identifies it as such.

7 Conclusions

We presented a supervised approach to classifying definitions as idiomatic or literal that more than dou-

Model	Rec.	Prec.	F1
Default	60.5	1	75.4
Annotated	78.3	76.7	77.5
Annotated+Default	89.2	79.0	**83.8**

Figure 7: Detection results.

bles the number of marked idioms in Wiktionary, even when training on incomplete data. When combined with the Lesk word sense algorithm, this approach provides a complete idiom detector for any phrase in the dictionary.

We expect that semi-supervised learning techniques could better recover the missing labels and boost overall performance. We also think it should be possible to scale the detection approach, perhaps with automatic dictionary definition discovery, and evaluate it on more varied sentence types.

Acknowledgments

The research was supported in part by the National Science Foundation (IIS-1115966) and a Mary Gates Research Scholarship. The authors thank Nicholas FitzGerald, Sarah Vieweg, and Mark Yatskar for helpful discussions and feedback.

References

J. Birke and A. Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

A. Budanitsky and G. Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

P. Cook, A. Fazly, and S. Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*.

P. Cook, A. Fazly, and S. Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the Language Resources and Evaluation Conference Workshop Towards a Shared Task for Multiword Expressions*.

M. Diab and P. Bhutada. 2009. Verb noun construction mwe token supervised classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*.

A. Fazly and S. Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

R. Fothergill and T. Baldwin. 2012. Combining resources for mwe-token classification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*.

Y. Freund and R.E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296.

M. Gedigian, J. Bryant, S. Narayanan, and B. Ciric. 2006. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*.

G. Katz and E. Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of Special Interest Group on the Design of Communication*.

L. Li and C. Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*. Springer.

E. Shutova, L. Sun, and A. Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the International Conference on Computational Linguistics*.

E. Shutova, S. Teufel, and A. Korhonen. 2012. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

T. Zesch, C. Müller, and I. Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the International Conference on Language Resources and Evaluation*.