Combining resources for MWE-token classification

Richard Fothergill and Timothy Baldwin
Department of Computing and Information Systems
The University of Melbourne
VIC 3010 Australia
r.fothergill@student.unimelb.edu.au, tb@ldwin.net

Abstract

We study the task of automatically disambiguating word combinations such as *jump the gun* which are ambiguous between a literal and MWE interpretation, focusing on the utility of type-level features from an MWE lexicon for the disambiguation task. To this end we combine gold-standard idiomaticity of tokens in the *OpenMWE* corpus with MWE-type-level information drawn from the recently-published *JDMWE* lexicon. We find that constituent modifiability in an MWE-type is more predictive of the idiomaticity of its tokens than other constituent characteristics such as semantic class or part of speech.

1 Introduction

A **multiword expression (MWE)** is a phrase or sequence of words which exhibits idiosyncratic behaviour (Sag et al., 2002; Baldwin and Kim, 2009). The nature of this idiosyncracy may be purely distributional — such as *hot and cold* being more common than *cold and hot* — but in this paper we study MWEs with idiosyncratic semantics. Specifically we are concerned with expressions such as *jump the gun* which are ambiguous between a literal interpretation of "to leap over a firearm", and an idiomatic interpretation of "to act prematurely".

While MWEs are increasingly entering the mainstream of NLP, the accurate identification of MWEs remains elusive for current methods, particularly in the absence of MWE type-specialised training data. This paper builds on the work of Hashimoto et al. (2006) and Fothergill and Baldwin (2011) in exploring whether type-level MWE properties sourced from an idiom dictionary can boost the accuracy of crosstype MWE-token classification. That is, we attempt to determine whether token occurrences of ambiguous expressions such as *Kim jumped the gun on this issue* are idiomatic or literal, based on: (a) annotated instances for MWEs other than *jump the gun* (e.g. we may only have token-level annotations for *kick the bucket* and *throw in the towel*), and (b) dictionary-based information on the syntactic properties of the idiom in question.

We find that constituent modifiability judgments extracted from the idiom dictionary are more predictive of the idiomaticity of tokens than other features of the idiom's constituents such as part of speech or lexeme. However, violations of the dictionary's modifiability rules have variable utility for machine learning classification, being suggestive of the literal class but not definitive. Finally, we present novel results illuminating the effectiveness of contextual semantic vectors at MWE-token classification.

2 Related Work

The *OpenMWE* corpus (Hashimoto and Kawahara, 2009) is a gold-standard corpus of over $100,000$ Japanese MWE-tokens covering 146 types. It is the largest resource we are aware of which has hand-annotated instances of MWEs which are ambiguous between a literal and idiomatic interpretation, and has been used by Hashimoto and Kawahara (2009) and Fothergill and Baldwin (2011) for supervised classification of MWE-tokens using features capturing lexico-syntactic variation and traditional semantic features borrowed from **word sense disambiguation (WSD)** . Similar work in other languages has been performed by Li and Sporleder (2010) and Diab and Bhutada (2009). We build on this work in exploring the use of MWE-type-level features drawn from an idiom dictionary for MWE identification.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 100–104,
Montréal, Canada, June 7-8, 2012.

Hashimoto and Kawahara (2009) developed a variety of features capturing lexico-syntactic variation but only one — a Boolean feature for "internal modification", which fired only when a non-constituent word appeared between constituent words in an MWE-token — had an appreciable impact on classification. However, they found that this effect was far overshadowed by semantic context features inspired by WSD. That is, treating each MWE-type as a word with two senses and performing sense disambiguation was far more successful than any features based on lexico-syntactic characteristics of idioms. Intuitively, we would expect that if we had access to a rich inventory of expression-specific type-level features encoding the ability of the expression to participate in different syntactic alternations, we should be better equipped to disambiguate token occurrences of that expression. Indeed, the work of Fazly et al. (2009) would appear to support this hypothesis, in that the authors used unsupervised methods to learn type-level preferences for a range of MWE types, and demonstrated that these could be successfully applied to a token-level disambiguation task.

Hashimoto and Kawahara (2009) trained individual classifiers for each MWE-type in their corpus and tested them only on instances of the type they were trained on. In contrast to this **type-specialised classification**, Fothergill and Baldwin (2011) trained classifiers on a subset of MWE-types and tested on instances of the remaining held-out MWE-types. The motivation for this **crosstype classification** was to test the use of data from the *OpenMWE* corpus for MWE-token classification of MWE-types with no gold-standard data available (which are by far the majority). Fothergill and Baldwin (2011) introduced features for crosstype classification which captured features of the MWE-type, reasoning that similar expressions would have similar propensity for idiomaticity. We introduce new MWE-type features expressing the modifiability of constituents based on information extracted from an MWE dictionary with wide coverage.

Fothergill and Baldwin (2011) expected that WSD features — however successful at type specialised classification — would lose their advantage in crosstype classification because of the lack of a common semantics between MWE-types. However, this turned out not to be the case, with by far the most successful results arising again from use of WSD features. This surprising result raises the possibility of distributional similarity between the contexts of idiomatic MWE-tokens of different MWE-types, however the result was not explained or explored further. In this paper we offer new insights into the distributional similarity hypothesis.

The recently-published *JDMWE* (*Japanese Dictionary of Multiword Expressions*) encodes type-level information on thousands of Japanese MWEs (Shudo et al., 2011). A subset of the dictionary has been released, and overlaps to some extent with the MWE-types in the *OpenMWE* corpus. *JDMWE* encodes information about lexico-syntactic variations allowed by each MWE-type it contains. For example, the expression *hana wo motaseru* — literally "to have [someone] hold flowers" but figuratively "to let [someone] take the credit" — has the syntactic form entry *[N wo] *V30*. The asterix indicates modifiability, telling us that the head [V]erb *motaseru* "cause to hold" allows modification by non-constituent dependents – such as adverbs – but the dependent [N]oun *hana* "flowers" does not.

3 Features for classification

We introduce features based on the lexico-syntactic flexibility constraints encoded in *JDMWE* and compare them with similar features from related work.

3.1 Type-level features

We extracted the modifiability flags from the syntactic field of entries in *JDMWE* and generated a feature for each modifiable constituent, identified by its position in the type's parse tree. The motivation for this is to allow machine learning algorithms to capture any similarities in idiomaticity between MWE-types with similar modifiability.

Fothergill and Baldwin (2011) also aimed to exploit crosstype similarity with their *type* features. They extracted lexical features (part-of-speech, lemma and semantic category) of the type headword and other constituents. We use these features as point of contrast.

3.2 Token features

An **internal modifier** is a dependent of a constituent which is not a constituent itself but divides an MWE-token into two parts, such as the word *seven* in *kick*

seven buckets. Features in related work have flagged the presence of any internal modifier unconditionally (Hashimoto and Kawahara, 2009; Fothergill and Baldwin, 2011). We introduce a refined feature which fires only when a MWE-token has an internal modifier which violates the constituent modification constraints encoded in *JDMWE*.

JDMWE modifiability constraints could also be construed to proscribe *external* modifiers. Sentential subjects and other external arguments of the head verb are too common to be sensibly proscribed but we did include a feature flagging proscribed external modification of leaf constituents such as *water* in *kick the bucket of water*. This feature effectively refines the **adnominal modification** feature of Hashimoto and Kawahara (2009) which indiscriminately flags external modifications on a leaf noun.

We include in our analysis a contrast of these features to token-based features in related work. The closest related features are those focussed on the MWE characteristic of lexico-syntactic fixedness termed **idiom** features by Hashimoto and Kawahara (2009) and Fothergill and Baldwin (2011):

- the flag for internal modification;
- the flag for adnominal modification;
- lexical features such as part-of-speech, lemma and semantic category extracted from an internal or adnominal modifier;
- inflections of the head constituent.

Additionally, we include WSD-inspired features used by Hashimoto and Kawahara (2009) and Fothergill and Baldwin (2011). These are all lexical features extracted from context, including part-of-speech, lemma and semantic category of words in the paragraph, local and syntactic contexts of the MWE-token. These features set the high water mark for classification accuracy in both type-specialised and crosstype classification scenarios.

3.3 Example JDMWE feature extraction

The following is a short literal token of the example type from Section 2, with numbered constituents: *kireina hanawo(2) motaseta(1)* ("[He] had [me] hold the pretty flowers"). The *JDMWE* features emitted for this token are the type feature *modifiable(1)* and the token feature *proscribed_premodifier(2)*.

4 Results

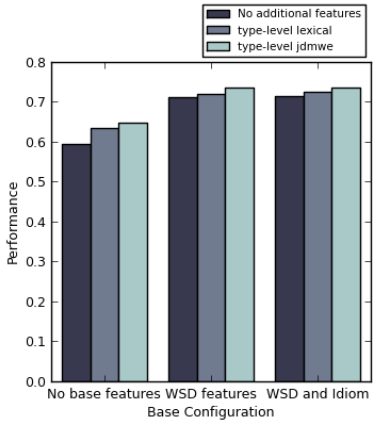
We worked with a subset of the *OpenMWE* corpus comprising those types having: (a) an entry in the released subset of the *JDMWE*, and (b) both literal and idiomatic classes represented by at least 50 MWE-tokens each in the corpus. This leaves only 27 MWE-types and $23,392$ MWE-tokens and means that our results are not directly comparable to those of Hashimoto and Kawahara (2009) and Fothergill and Baldwin (2011). The release of the full *JDMWE* should enable more comparable results.

We constructed a crosstype classification task by ten-fold cross validation of the MWE-types in the *OpenMWE* subset, with micro-averaged results. Training sets were the union of all MWE-tokens of MWE-types in a partition. The majority class was the idiomatic sense and provided a baseline accuracy of 0.594. *Support Vector Machine* models with linear kernels were trained on various feature combinations using the *libSVM* package.

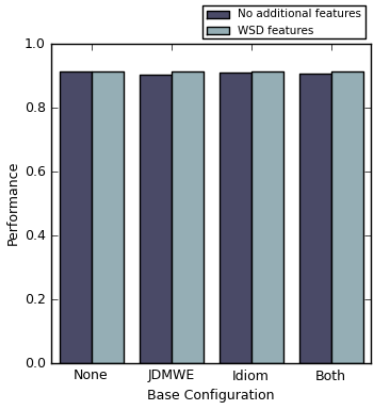
Our *JDMWE* type-level features performed comparatively well at the crosstype task, with an accuracy of 0.647, at 5.3 percentage points above the baseline. This is a marked improvement on the lexical type-level features from related work, which achieved an accuracy of 4.0 points above baseline. As has been observed in related work, the accuracy gained by using type-level features is much smaller than the token-level WSD features. However, the relative performance of the *JDMWE* type features to the lexical type features is sustained in combination with other feature types, as shown in Figure 1a.

Our *JDMWE* token-level features on the other hand perform quite badly at crosstype classification. When measured against the baseline or used to augment other token features, they degraded or only marginally improved performance. The fact that using these features resulted in worse-than-baseline performance suggests that the constituent modifiability features extracted from *JDMWE* may not be strict constraints as they are construed.

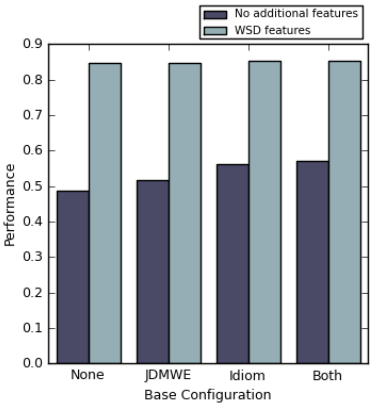
To better examine the quality of the *JDMWE* constituent modifiability constraint features, we constructed a heuristic classifier. The classifier applies the idiomatic class by default, but the literal class to any MWE-token which violates the *JDMWE* constituent modifiability constraints. This classifier's



(a) Accuracy using *JDMWE* type-level features and lexical type-level features in combination with various token-level features



(b) Recall for idiomatic instances for various feature combinations with and without WSD context features, in a type-specialised classification setting



(c) Recall for literal instances for various feature combinations with and without WSD context features, in a type-specialised classification setting.

Figure 1: Results

precision on the literal class was 0.624, meaning that fully 0.376 of modifiability constraint violations in the corpus occured for idiomatic tokens.

However, the classifier was correct in its literal class labels more than half the time so it achieved a better accuracy than the majority class classifer, at 0.612. As such, the heuristic classifier comfortably outperformed the *Support Vector Machine* classifier based on the same features. This shows that our poor results with regards to the *JDMWE* constraint violation features are due mainly to failures of the machine learning model to take advantage of them.

As to the strength of the constraints encoded in *JDMWE*, we found that 4.4% of all idiomatic tokens in the corpus violated constituent modification constraints, and 10.8% of literal tokens. Thus the constraints seem sound but not as rigid as presented by the *JDMWE* developers.

Figure 1a shows that even with our improvements to type-level features, the finding of Fothergill and Baldwin (2011) that WSD context features perform best at crosstype classification still holds. We cannot fully account for this, but one observation regarding the results of our type-specialised evaluation may have bearing on the crosstype scenario.

For our type-specialised classification task we performed cross-validation for each MWE-type in isolation, aggregating final results. Some types had a literal majority class, so the baseline accuracy was 0.741. Figure 1b shows that type-specialised classification performance is basically constant when restricting analysis to only the idiomatic test instances. The huge performance boost produced through the use of WSD features occurs only on literal instances (see Figure 1c). That is, our type-specialised classifiers are capturing distributional similarity of context for the literal instances of a MWE-type but not for the idiomatic instances. Since the contexts of idiomatic instances of the same MWE-type do not exhibit a usable distributional similarity, it is unlikely that crosstype similarities between *idiomatic* MWE-token contexts can explain the efficacy of WSD features for crosstype classification.

5 Conclusion

Using a MWE dictionary as input to a supervised crosstype MWE-token classification task we have shown that the constituents' modifiability characteristics tell more about idiomaticity than their lexical characteristics. We found that the constituent modification constraints in *JDMWE* are not hard-and-fast rules but do show up statistically in the *OpenMWE* corpus. Finally, we found that distributional similarity of the contexts of idiomatic MWE-tokens is unlikely to be the source of the success of WSD features on MWE-token classification accuracy.

References

Timothy Baldwin and Su Nam Kim. 2009. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Boca Raton, USA, 2nd edition.

Mona T. Diab and Pravin Bhutada. 2009. Verb noun construction MWE token supervised classification. In *MWE '09: Proceedings of the Workshop on Multiword Expressions*, pages 17–22, Singapore.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Richard Fothergill and Timothy Baldwin. 2011. Fleshing it out: A supervised approach to MWE-token and MWE-type classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand.

Chikara Hashimoto and Daisuke Kawahara. 2009. Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation*, 43:355–384.

Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006. Detecting Japanese idioms with a linguistically rich dictionary. *Language Resources and Evaluation*, 40:243–252.

Linlin Li and Caroline Sporleder. 2010. Linguistic cues for distinguishing literal and non-literal usages. In *Coling 2010: Posters*, pages 683–691, Beijing, China.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 189–206, Mexico City, Mexico.

Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe. 2011. A comprehensive dictionary of multiword expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, USA.