Fleshing it out: A Supervised Approach to MWE-token and MWE-type Classification

Richard Fothergill and Timothy Baldwin
Department of Computer Science and Software Engineering
The University of Melbourne
VIC 3010 Australia
r.fothergill@student.unimelb.edu.au, tb@ldwin.net

Abstract

Although some multiword expressions (MWEs) like *How do you do?* have exclusively idiomatic meaning, other MWE-types like the phrase *kick the bucket* may be idiomatic or literal depending on context. The recently developed *OpenMWE* corpus provides the largest freely available collection of annotated MWE-tokens suitable for supervised classification, but so far its potential has only been superficially investigated and only for classification of MWE-types in the corpus. Instead, we train and evaluate classifiers for crosstype classification and introduce novel features specialised to this task. Our best crosstype classifiers performed as well on non-trained MWE-types as a majority class baseline which has knowledge of the MWE-type.

1 Introduction

A **multiword expression (MWE)** is an idiosyncratically interpreted linguistic unit which consists of more than a single word (or "crosses word boundaries") (Sag et al., 2002; Baldwin and Kim, 2009). The nature of these idiosyncrasies can vary greatly — from *traffic light* and *street light* which are remarkable only in that they are not interchangeable — to *how do you do?*, the meaning of which is non-compositional in modern English.

MWEs will typically resist lexico-syntactic variation to some extent (Fazly et al., 2009). For example, the phrase *a picture is worth a thousand words* does not allow the freedom of lexical substitution and modification that its constituent words would usually enjoy, making otherwise equivalent variations — *an image is worth fifty score words*, *a picture is worth approximately a thousand words* — sound wrong or at least unnatural to a native speaker.

The meaning of a MWE as a whole may not derive literally from the composition of its constituent words (Baldwin et al., 2003). For example, the English expression *kick the bucket* may be used to refer to *death* in any number of ways unrelated to either kicking or buckets. In these cases we say the MWE has an idiomatic interpretation.

When talking about MWEs we make a distinction between an MWE lexeme out of context, which we call an **MWE-type**, and specific instances of the MWE in context, which we call an **MWE-token**.

Some MWE-types have only an idiomatic meaning, such as the English greeting *How do you do?*, interpretation of which can be perplexing if attempted literally. However for others, the literal uses are still perfectly valid, and individual MWE-tokens may be ambiguous between an idiomatic and literal meaning. For example, *kick the bucket* may indeed refer to a violent act against a bucket and have nothing to do with death at all.

Relation to Word Sense Disambiguation

MWE-token disambiguation can be approached as if it were a word sense disambiguation (WSD) task where the MWE-types correspond to word types and MWE-tokens to word tokens (Hashimoto and Kawahara, 2009). In this conception, the analogue of word senses are the idiomatic and literal classes for an MWE-type.

In WSD, supervised methods are by far the most successful but large amounts of data are required for *each* word type to be disambiguated (Navigli, 2009). However, unlike in WSD, we can expect to

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 911–919,
Chiang Mai, Thailand, November 8 – 13, 2011.

find *some* linguistic commonality between the idiomatic senses of distinct MWE-types. This leads us to hope for more general **crosstype** classification algorithms, which might alleviate the knowledge acquisition bottleneck. In this paper we will refer to WSD (more specifically "word expert") style classification of ambiguous MWE-tokens as **type-specialised** classification.

The **first sense** or **majority class** baseline, in which a word is always labelled with its predominant sense, is known to perform very well at WSD due to a Zipfian distribution of senses (Preiss et al., 2009). We expect no different for the MWE-token disambiguation task where the two-class problem is virtually guaranteed a majority class baseline accuracy of over 50%. There has been work in unsupervised first sense learning for WSD using lexical resources (McCarthy et al., 2007), but the de facto baseline in WSD is a supervised first sense baseline (Navigli, 2009). We make use of a **type-specialised baseline**, which is a supervised majority class baseline modelled on the WSD first sense baseline. We also introduce a **corpus baseline** which, calculated based on idiomatic and literal counts of a collection of MWE-types, is the type-specialised baseline's crosstype analogue.

Our Contribution

In this paper we explore the supervised classification of ambiguous MWE-tokens using the *OpenMWE* corpus of Japanese idioms (Hashimoto and Kawahara, 2009). We introduce new features tailored to the crosstype classification task and refine features for type-specialised classification. To our knowledge, our experiments are the largest in supervised MWE-token classification to date. We explore more deeply the interaction between several aspects of the task, including differences between crosstype and type-specialised classification; combinations of major classes of features; and finally, the size of the training corpus.

We find that:

1. Our new WSD inspired features offer consistent improvements in performance, and our new idiom features usually offer marginal improvements. The extended WSD and idiom features used together for type-specialised classification yield state-of-the-art results in terms of raw performance.

2. Our new features for crosstype classification interact with the task and with other features in interesting ways, and in some cases give substantial improvements to performance.

3. Our best results in crosstype classification use *only* our extended WSD features. Despite using no tagged data for the target MWE-types, they achieved a performance in excess of the (supervised) type-specialised baseline.

This last result is significant because it demonstrates the readiness of our crosstype classifiers to work on previously unseen MWE-types. It is also interesting because it uses only semantic features and none of the lexico-syntactic fixedness features widely expected to be effective for MWE classification (Fazly et al., 2009; Hashimoto et al., 2006; Li and Sporleder, 2010).

2 Related Work

The *OpenMWE* corpus was compiled by Hashimoto and Kawahara (2009). To our knowledge it is by far the largest freely available gold-standard corpus of ambiguous MWE-tokens in any language, comprising 146 ambiguous MWE-types with 102,856 annotated MWE-tokens in total.

Apart from the enormous task of constructing the *OpenMWE* corpus, Hashimoto and Kawahara (2009) also used it to perform some experiments with supervised MWE-token classification. Noting the similarities between this task and WSD, they employed the most effective WSD features and machine learning algorithm surveyed by Lee and Ng (2002). They also included linguistic features explored by Hashimoto et al. (2006) designed to capture the relative fixedness of Japanese idioms, which we will refer to as idiom features. The machine learning algorithm used was *Support Vector Machines* and models were trained on the WSD features with various combinations of the idiom features. Type-specialised classifiers were trained for the 90 MWE-types which were deemed to have sufficient idiomatic and literal examples in the corpus. The model trained on WSD features was found to improve greatly on the type-specialised baseline, with some additional performance added by one of the idiom features.

Only being able to classify MWE-tokens of the 90 MWEs with sufficient training examples is a severe limitation on the usefulness of type-specialised classifiers in natural language processing applications. Our goal is to escape this limita-

tion by training classifiers which work on MWE-types on which they have not been trained. To that end, we have extended the features used by Hashimoto and Kawahara (2009) with complementary features and introduced a new class of features designed for crosstype classification.

Li and Sporleder (2010) conducted a thorough investigation of features used for supervised MWE-token classification. Context features similar to the WSD features of Hashimoto and Kawahara (2009) were used, as were a number of linguistically motivated features. Like us, Li and Sporleder (2010) performed crosstype classification. Unfortunately many of the features were too sparse to have a significant effect and only the context features produced significant results. This may have been due to the relatively small size of the corpus used, which comprised around 4000 MWE-tokens across 13 MWE-types. In our results the context features still dominate, but the effects of idiom based features can be seen and those features hold up well on their own as well.

Diab and Bhutada (2009) described a novel supervised MWE-token classification system based on a sequence labelling model. Unlike our method, their model identifies the position of the token in the text as part of the process. Like Li and Sporleder (2010), the size of the corpus is small (2500 MWE-tokens of 53 MWE-types), and classifiers were trained on collections of MWE-types. Anecdotally, the classifiers were able to pick some MWEs out of running text without even knowing their constituents beforehand, however their performance at this was not tested. A major finding of Diab and Bhutada (2009) was that reducing the feature space of context features by replacing word lemmas with their named-entity category had a significant positive effect on classification performance, a finding that is consolidated by Diab and Krishna (2009).

Fazly et al. (2009) observed that idiomatic uses of MWEs tend to occur in one of a small set of canonical forms, and developed an unsupervised method for learning these canonical forms, based on a set of linguistically-motivated features which built on the work of Cook et al. (2007). They applied the learned set of canonical forms to the task of MWE-token identification with remarkable success. Our method similarly uses linguistically-motivated features to perform MWE-token identification, but using a cross-type supervised model.

3 Feature Extraction

We extracted features in three main groups: WSD features, idiom (token) features and idiom type features. The first two categories include the features used by Hashimoto and Kawahara (2009) and our own extensions. The third category was introduced by us and is specialised to crosstype classification.

3.1 Preprocessing

Our feature extraction makes extensive use of the Japanese dependency parser *KNP* (Kurohashi and Nagao, 1994),[1] however features should be reproducible in other languages with a suitable dependency parser, morphological analyser, and electronic thesaurus or ontology. Each instance in the corpus was preprocessed by running it through *KNP* to extract specific linguistic information.[2] To help elucidate both the details of our feature extraction and how it might be replicated for another language, we will look at the information extracted for the sentence in Example (1):

(1) 桂子さんは、サッカーの腕を
keiko-saN-wa, sakkā-no ude-o
Keiko-TOPIC soccer-GEN arm-OBJ
上げた。
ageta.
raised.

“Keiko raised her soccer arm.” (literal)

“Keiko improved her skills at soccer.” (idiomatic)

Japanese is a non-segmenting language in that it has no clearly marked word boundaries. In Figure 1, Example (1) has been segmented by *KNP* into tokens[3] which are then grouped into chunks. Each chunk has a *parent* link to a higher chunk describing the dependency parse of the sentence. In the conventional word order for Japanese, dependency links are always forwards so the head of a phrase is also its rightmost (final) chunk. From Figure 1, we see that *keiko* “Keiko” and *ude* “arm” are the dependents of *ageta* “raised”, with *sakkā* “soccer” modifying *ude* “arm”. This gives rise

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/?KNP

[2] *KNP*’s memory usage is high and on some sentences exceeded the available memory in our machine (32GB). Those instances were excluded from our analysis.

[3] In fact, *KNP* delegates the initial segmentation and token feature annotation to the morphological analyser *Juman*.

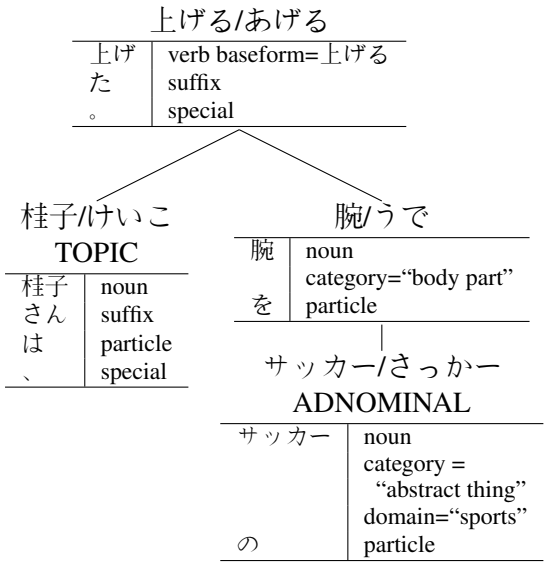

Figure 1: English summary of the select parts of the output of *KNP* when run on the sentence in Example (1)

to the incorrect literal interpretation *#Keiko raised her soccer arm*. In fact, *ude-o ageru* "to improve one's skills" is an ambiguous MWE.

In this paper when we refer to *words* we are in fact referring to the chunks returned by *KNP*. This is the most appropriate level of segmentation for our purposes because it is the level at which the dependency parse exists and because the tokenisation level includes affixes and particles which would need filtering. *KNP* labels each chunk with a normalised form, which we use as the *lemma* of the word for our feature extraction. We also extract the part of speech, category and domain of tokens corresponding most closely to the lemma, as additional features of the word.

The category and domain information performs a similar function to the named entity information used by Diab and Bhutada (2009): it collapses classes of words into a single feature while retaining relevant semantic information. An information source such as an ontology or thesaurus could be substituted for use in other languages. Hashimoto and Kawahara (2009) translate the *category* output of *KNP* as *hypernym*, and we will adopt the same terminology hereafter.

Returning to Figure 1, take note of the *TOPIC* flag on the first chunk and the ADNOMINAL flag on the second. *KNP* produces many chunk annotations; we made use of five main kinds:

ADNOMINAL appearing on adnominal modifiers;

TOPIC appearing on sentential topics and emphasised chunks;

VOICE of inflected verbs;

NEGATED denoting negation; and

VOLITIONAL denoting a volitional modality.

For all but the volitional modality group, *KNP* outputs only one or two annotation variants (e.g. there are two voices in Japanese: passive and causative). For the volitional modality, Hashimoto and Kawahara (2009) used five classes: request, invitation, order, volition and prohibition. We made use of the same subset of modalities output by *KNP*.

The *KNP* chunk annotations we used chiefly capture inflections on words in the text, something which Diab and Bhutada (2009) approximate with a character n-gram feature. For implementation in other languages, the n-gram heuristic or any available morphological analyser might be used.

3.2 Word sense disambiguation features

We adopted the WSD inspired features of Hashimoto and Kawahara (2009) and extended them with our own new features.

Hashimoto and Kawahara (2009) features

Hashimoto and Kawahara (2009) put together a set of features based on WSD best practice as espoused by Lee and Ng (2002). They included:

1. full context, in the form of bag features for lemma, hypernym and domain of single words in the full surrounding context.[4]

2. local context, in the form of part of speech and lemma features for indexed word offsets to each side of the MWE.

3. syntactic context, in the form of lemmas and POS for context words in a syntactic relationship with either the first or last constituent word of the MWE.

For more details on implementation of these features, see Hashimoto and Kawahara (2009).

[4]Note that due to the way the corpus was constructed, the full context is a single sentence, albeit a long one.

full context features

lemma keiko, majime, ...
hypernym abstract thing, body part, person.
domain sports, familial associations.

local context features

lemma majime$_{-3}$, reNshū$_{-2}$, sakkā$_{-1}$, ...
(majime, sakkā)$_{-3,-1}$, ...
POS adjective$_{-3}$, noun$_{-2}$, noun$_{-1}$, ...
(adjective, noun)$_{-3,-1}$, ...
hypernym abstract thing$_{-2}$, abstract thing$_{-1}$, ...
(NULL, abstract thing)$_{-3,-1}$, ...
domain sports$_{-1}$, ...
(NULL, sports)$_{-3,-1}$, ...

syntactic context features

lemma sakkā$_{\text{child}}$
POS noun$_{\text{child}}$
hypernym abstract thing$_{\text{child}}$
domain sports$_{\text{child}}$

Figure 2: A sample of the WSD inspired features extracted for Example (2).

New WSD features

We introduce hypernym and domain features for the local and syntactic contexts. Use of hypernym and domain features in the syntactic context is particularly interesting. The intent of the syntactic features is to capture selectional restrictions involving constituents of the MWE. Violation of selectional restrictions for or by constituents of the MWE leads us to strongly suspect an idiomatic usage. In the case of Example (2) we see that having *sakkā* "soccer" modifying *ude* "arm" is strongly indicative of the idiomatic *ude-o ageru* "to raise skills". For this MWE, any sport has the same implication. We can see from Figure 1 that *KNP* has extracted the domain *sports* for *sakkā* "soccer" so a classification algorithm can use this feature to make a valid generalisation. Our local context hypernym and domain features are less targeted, but we consider them to be worthwhile in light of the success of named-entity features in the literature (Diab and Bhutada, 2009).

Note that we use the same definitions of local and syntactic context as Lee and Ng (2002), with the specialisations to MWEs outlined by Hashimoto and Kawahara (2009).

Figure 2 contains a sample of all original and new WSD features, extracted for the MWE-token in Example (2), which expands Example (1):

(2) 桂子さんは　真面目に 練習して、
keiko-saN-wa majime-ni reNshū-shite,
Keiko-TOPIC diligent-ly practice,
サッカーの 腕を　　上げた。
sakkā-no ude-o ageta.
soccer-GEN arm-OBJ raised.
お母さんは　喜んだ。
okāsan-wa yorokoNda.
Mother-TOPIC pleased.
"Keiko practised diligently and improved her skills at soccer. Her mother was pleased."

3.3 Idiom token features

As with the WSD features, we adopted the idiom features of Hashimoto and Kawahara (2009) with additional features of our own.

Hashimoto and Kawahara (2009) features

The idiom features of Hashimoto and Kawahara (2009) included a single binary feature for each of the *KNP* chunk annotation groups listed at the end of Section 3.1. For each of the groups, the feature fires if one of the annotations appears on a relevant word in the MWE. Details of each are given in the outline of our extensions below.

The final idiom feature of Hashimoto and Kawahara (2009) was the adjacency feature. This feature fires if the constituents of the MWE are contiguous, i.e., there are no intervening chunks. They found that this feature had a greater impact on classification performance than the other idiom features combined.

New idiom features

Each of the boolean idiom features of Hashimoto and Kawahara (2009) captures some variation of the form of the MWE. We introduce features capturing details on the kind of variation:

- The ADNOMINAL modification feature fires if a non-constituent adnominal modifies a noun in the MWE. We include features for the lemma, POS, hypernym and domain of such a modifier whenever it exists.
- The TOPIC feature fires if a constituent noun is marked as a sentential topic. In this case we include features for the lemma, POS, hypernym and domain of the constituent.
- The VOICE, NEGATION and VOLITIONAL features fire if the MWE has a head

verb and it has voice marking or is negated. We include features specifying what for of voice, negation or volitional marking is used.

- Finally for the adjacency feature, we include lemma, POS, hypernym and domain features from a single intervening chunk where any existed. When more than one is found, we take the rightmost.

In the sentence of Example (2), features only fire for adnominal modification. Since *sakkā* "soccer" modifies the constituent *ude* "arm", the adnominal modification boolean feature fires, as do our lemma, POS, hypernym and domain features for the modifier: *sakkā*, *noun*, *abstract thing* and *sports* respectively. As was the case when *sakkā* "soccer" was considered in its role as a modifier of *ude* "arm", we note that it is in fact informative that the adnominal modifier is a sport and not, for example, a person.

3.4 Type features

The features we have discussed so far have, for the most part, ignored the constituent words of the MWE-type itself. For type-specialised classifiers this is inconsequential since the constituents are constant. However features of the MWE-type may be important for crosstype classification where similarities between different MWEs could be leveraged. Therefore, for each MWE-type, we use the lemma, POS, hypernym and domain of the headword in particular and a bag feature for each across all words in the MWE.

One motivation for these features is to allow a crosstype classifier to make more informed decisions when confronted with tokens of types it was trained on. For example, if a collection of constituents has been encountered in training, a supervised statistical classifier may capture the prior probability for idiomaticity of the training MWE-type. For crosstype classification, some constituents — in particular the headword — may be indicative of the relative idiomaticity of an idiom. For example, common verbs such as *take* and *make* are common in idioms such as *take a shot* and *make a stand*.

4 Results

We evaluated classifiers using combinations of the three main classes of features. For all tasks, a ten-fold cross-validation partitioning was used, and a feature count cutoff of one was used to filter out uninformative features. Given the binary classification nature of the task we used accuracy as our performance metric, microaveraging across all instances in the corpus.

Feature Types			Accuracy	
idiom	*type*	*wsd*	*basic*	*extended*
	∗		0.623	–
∗			0.630	0.627
∗	∗		0.626	0.651
		∗	0.737	0.745
	∗	∗	0.736	0.743
∗		∗	0.738	0.746
∗	∗	∗	0.739	0.745
corpus baseline			0.612	–
type-specialised baseline			0.741	–

Table 1: Results of combining different features for crosstype classification. "Basic" results restrict the idiom and WSD features to those of Hashimoto and Kawahara (2009); "extended" results include our extensions.

For statistical significance we used the sign test because our crosstype classification testing partitions were unevenly weighted, making a t-test inappropriate. Unless otherwise stated, comparisons were significant with $p < 0.05$.

We constructed two kinds of majority class baseline using class counts from the corpus: the corpus baseline, which achieved an accuracy of 0.612, and the type-specialised baseline, with an accuracy of 0.741. All other systems were linear kernel *Support Vector Machine* models trained using the *libSVM* package.[5]

4.1 Crosstype classification

For the purposes of testing crosstype classification we partitioned the set of 90 MWE-types for cross-validation. Thus classifiers were trained on the instances of 81 types and tested on the instances of the 9 unseen types. Note that since the corpus contains a different number of instances for each type, the partition size was not strictly constant. Results across all feature combinations appear in Table 1.

The idiom type and token features did manage to improve on the corpus baseline by a little over one percentage point each. However, this is over

[5]We initially used the *TinySVM* package and quadratic kernels of Hashimoto and Kawahara (2009) for comparability reasons, but eventually changed system and kernel for consistency and speed of convergence.

Feature Types			Accuracy	
idiom	*type*	*wsd*	*basic*	*extended*
	∗		0.741	–
∗	∗		0.748	0.752
∗			0.630	0.639
		∗	0.844	0.847
	∗	∗	0.851	0.854
∗		∗	0.847	0.849
∗	∗	∗	0.854	0.856
corpus baseline			0.612	–
type-specialised baseline			0.741	–

Table 2: Results for classification of MWE-tokens of MWE-types seen in the training corpus. As in Table 1, "Basic" results restrict the idiom and WSD features to those of Hashimoto and Kawahara (2009).

ten percentage points behind the results when using WSD features alone. In fact, the WSD features *with our extensions* achieved effectively our best results. The addition of idiom features with our extensions achieved fractionally better performance without statistical significance and all feature combinations which did not include our full complement of WSD features had lower performance. The WSD features' performance exceeds even the type-specialised baseline, which was built on gold-standard data which the crosstype classifier has no access to.

It is a surprising result, for two reasons: first, that idiom features are widely assumed to be a key information source, particularly for unsupervised disambiguation (Cook et al., 2007), and second, that WSD features — paragraph context in particular — are typically used as a model of the *meaning* of a token. It is counterintuitive that models of semantics are more informative than models of lexico-syntactic variations when the testing and training sets that are explicitly disjoint with respect to MWE-type.

The idiom token or type features alone did not stand up well in comparison. However, we note that combining our type features with the complete idiom token features provided a disproportionate boost to a classification accuracy of almost four percentage points above the baseline.

What happens when a nominally crosstype classifier encounters an instance of a MWE-type which it *has* seen in its training set? To test this, we partitioned the corpus stratified across types.

Feature Types		Accuracy	
idiom	*wsd*	*basic*	*extended*
∗		0.768	0.769
	∗	0.882	0.886
∗	∗	0.886	0.890
type-specialised baseline		0.741	–

Table 3: Results of combining different features for type-specialised classifiers. "Basic" results restrict the idiom and WSD features to those of Hashimoto and Kawahara (2009); "extended" results include our extensions.

That is, classifiers were trained on 90% of instances from *all* MWE-types in the corpus and tested on the remaining 10%. The results are shown in Table 2.

The idiom features once again improved a couple of percentage points on the baseline. In this instance, the type features produced much better results, achieving the same results as the type-specialised majority class baseline. It is not possible for any deterministic classifier to do better on the same input because the idiom type features are constant across all instances of a MWE-type.

Once again the WSD features did far better than any of the others at 23 percentage points over the corpus baseline and over ten points above even the type-specialised baseline. This is more to be expected than the equivalent result for crosstype classification since, by their origin, WSD features are designed to capture differences in semantics for known types.

The indisputable dominance of WSD features observed in these experiments warrants further investigation, which we leave for future work.

4.2 Type-specialised classification

If a known MWE-type can be explicitly detected, the general crosstype classifier need not be used: we can fall back on type-specialised classifiers like those of Hashimoto and Kawahara (2009). To see how much better we can do by selecting an appropriate type-specialised classifier, we trained and tested classifiers on the same partitioning as the previous task but restricted to instances of one MWE-type at a time. The results appear in Table 3.

In this case the idiom features improved a little even on the type-specialised baseline. This indicates that the idiom features do contain information about MWE-token idiomaticity even if it does

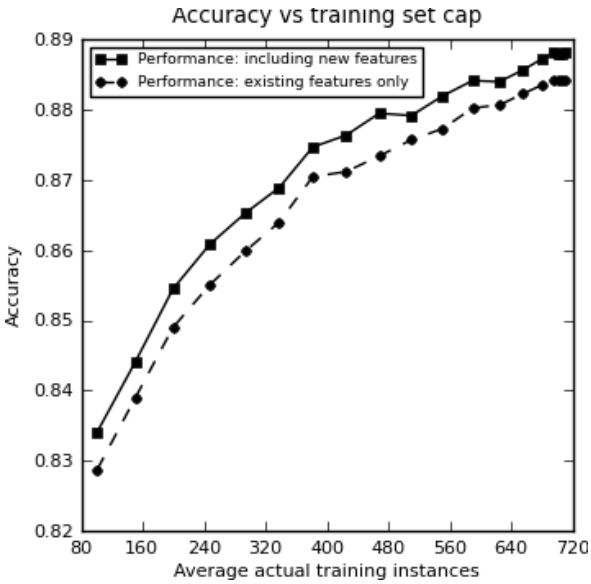

Figure 3: Results for type-specialised classifiers with capped training set sizes.

not generalise well across types.

The WSD features achieved close to the best results seen across all our experiments. An improvement of four percentage points is seen compared to the previous task, which is indicative of the noise introduced by simultaneously training on data of 90 idioms. Our best results were achieved using all of the idiom and WSD features together, hitting an even 0.890.

Finally, we used the type-specialised task to investigate the significance of the size of the *OpenMWE* corpus. To do this we measured cross-validation accuracy while limiting the total number of training instances. Training set size caps ranging between 100 and 1000 instances were used, but in practice most of the MWE-types had between 500 and 1000 available training instances, so the average actual training instances used was less than the cap. We performed these experiments using our complete WSD and idiom features and, for comparison, with the original features of Hashimoto and Kawahara (2009). The results appear in Figure 3.

Even with 100 instances per MWE-type, we achieved an accuracy of 0.834, which is an appreciable improvement on the type-specialised baseline. However the data show a definite positive trend with the number of instances, reaching 0.884 under a cap of 650 instances (and 589 average actual instances) per MWE-type.

Setting the maximum number of instances per MWE-type to 1000 achieved an accuracy of 0.888. Additionally, when restricted to the original features used by Hashimoto and Kawahara (2009) a performance of 0.884 is observed. Since Hashimoto and Kawahara (2009) also capped instance counts at 1000 this is our most comparable result to their best of 0.893. We note that with our new features, results were consistently around half a percentage point higher, so consider this to be state of the art performance.

5 Conclusions

We have shown that crosstype classification of ambiguous MWE-tokens can surpass the type-specialised baseline while alleviating the requirement on labelled token instances, thus enabling classification of tokens of previously unseen MWE-types.

Our type features and new idiom features, working in concert with the idiom features of Hashimoto and Kawahara (2009), substantially increase crosstype classification performance over the baseline. However their effect is wholly subsumed by the inclusion of WSD features. On type-specialised classification, our new idiom and WSD features achieve more consistent gains.

Finally, we conclude that the size of the *OpenMWE* corpus raises potential performance by leaps and bounds, but additional performance is still to be had by more data.

For future work we would like to investigate the dominance of WSD features at crosstype classification. The success of semantic features where the training and test sets have — by design — different semantics, making this an intriguing counter-intuitive result, as does the relatively poor performance of features targeted at linguistic properties of MWEs.

Acknowledgements

This research was supported by Australian Research Council grant no. DP0988242.

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.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model

of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Paul Cook, Afsaneh Fazly, and Suzanne 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*, pages 41–48, Prague, Czech Republic.

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.

Mona T. Diab and Madhav Krishna. 2009. Handling sparsity for verb noun MWE token classification. In *GEMS '09: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 96–103, Athens, Greece.

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

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.

Sadao Kurohashi and Makoto Nagao. 1994. KN parser: Japanese dependency/case structure analyzer. In *Proceedings of the Workshop on Sharable Natural Language Resources*, Nara, Japan.

Yoong Keok Lee and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *EMNLP '02: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 41–48, Philadelphia, USA.

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

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2).

Judita Preiss, Jon Dehdari, Josh King, and Dennis Mehay. 2009. Refining the most frequent sense baseline. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 10–18, Boulder, USA.

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.