Automagically Inferring the Source Words of Lexical Blends

Paul Cook
Department of Computer Science
University of Toronto
Toronto, Canada
pcook@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
Toronto, Canada
suzanne@cs.toronto.edu

Abstract

Lexical blending is a highly productive and frequent process by which new words enter a language. A blend is formed when two or more source words are combined, with at least one them shortened, as in *brunch* (*breakfast*+*lunch*). We use linguistic and cognitive aspects of this process to motivate a computational treatment of neologisms formed by blending. We propose statistical features that can indicate the source words of a blend, and whether an unknown word was formed by blending. We present computational experiments that show the usefulness in these tasks of features tapping into the recognizability of the source words in the blend, in combination with their semantic properties.

1 Lexical Blends

Many natural language processing tasks depend on lexical resources that provide usage information for the words a system must handle. Since words can change in their usage over time, much research in computational linguistics has investigated the automatic learning of lexical information from existing resources, such as corpora, in order to keep lexicons accurate and up-to-date. Most lexical acquisition work has focused on learning syntactic and coarse-grained semantic properties of known words (e.g., Riloff and Jones, 1999; McCarthy, 2001; Korhonen, 2002). However, neologisms—newly created words, which enter the language on a regular basis—present a serious challenge for such efforts, since, by definition, we are not simply adding to or updating our existing knowledge about a word.[1] While available usage data may give strong clues as to the syntactic properties of neologisms, automatically determining their semantics poses a real obstacle to maintaining wide-coverage lexical resources.

Fortunately, people rarely create completely new words that give no clue as to their meaning. One common means for introducing novel words is subtractive word formation processes, in which existing words are clipped and possibly combined. Such methods include shortening (*lab* for *laboratory*), forming an acronym (*CL* for *computational linguistics*), or lexical blending (e.g., Algeo, 1977; Kreidler, 1979). In blending, two or more source words are combined to form a novel word—a blend—which captures some aspect of the semantics of both source words. Blending accounts for roughly 5% of the neologisms in the list of new words accumulated by Algeo (1991). In our study, we focus on two-word sequential blends, such as *brunch* (*breakfast* + *lunch*) and *ebonics* (*ebony* + *phonics*), which are very common type of blend. In these, a prefix of the first source word precedes a suffix of the second.[2]

Many existing English blends, such as *motel* (*motor* + *hotel*) and *meld* (*melt* + *weld*), would be included in a typical computational lexicon. Indeed, some such blends have become conventionalized to the point where some speakers do not even recognize them as blends; however, new blends are continually entering the language, and at an apparently increasing rate (Lehrer, 2003). Blending is used extensively to create novel proper nouns, especially the names of celebrities (*Brangelina*, *Brad Pitt* + *Angelina Jolie*) and companies or organizations (*Fruitopia*, *fruit* + *utopia*, and *Wikipedia*, *wiki* + *encyclopedia*). However, common nouns and other parts of speech are also frequently coined using blending;

[1] Here we do not consider the use of an existing word in a new sense or part of speech as a true neologism, since such usages inherit much from the prior usage(s) of the word.

[2] It is rare for blends to be non-sequential, as in *burble* (*bubble* + *murmur* (Algeo, 1977)), or composed of more than two source words, as in *turducken* (*turkey* + *duck* + *chicken*).

recent creations include *mathlete* (*math* + *athlete*), *chillax* (*chill* + *relax*), and *badong* (*bad* + *wrong*).

Automatically learning the meaning of novel lexical blends involves a sequence of challenging steps. First, given an unknown word, it must be decided whether the word is a blend. Second, once the word has been identified as a blend, its source words must be determined. Finally, the sense contributed by each source word, and the semantic relationship between them, must be identified. To our knowledge, none of these three steps has been addressed in the CL literature. In this study, we focus primarily on the second step of this process, that is, the problem of automatically determining the source words of a word that has been identified as a blend. We then present some preliminary results which indicate that our methods for this step can also form the basis for the first step of deciding whether an unknown word is indeed a blend.

2 Computational Models of Blends

Given an unknown word that we know is a blend, the goal is to identify the source words from which it is formed. Currently we assume that we are given no usage information for the word; we consider the possible impact of this assumption when we discuss our experimental results.

Our approach has two steps. First, we identify the set of all pairs of words which, based on orthography, could have formed the blend. We refer to this set as the *candidate set*, and the word pairs it contains as the *candidate pairs*. Then for each candidate pair we calculate a number of statistical features which are expected to be higher for the correct candidate pair and lower for others.

2.1 Finding Candidate Pairs

To create the candidate set for a given blend, we begin by considering each way of splitting the blend into a prefix and suffix.[3] We require that both the prefix and suffix be of length at least two to prevent the candidate sets from becoming excessively large. For example, for the blend *boatel* (*boat* + *hotel*) we consider each of the following prefix–suffix pairs: *bo, atel*; *boa, tel*; *boat, el*.

For each prefix–suffix pair, we extract from a dictionary all words beginning with that prefix and all words ending with that suffix. The Cartesian product of the two sets of words yields a set of candidate pairs. The candidate pairs for all prefix–suffix pairs are combined to form the full candidate set for a blend.

2.2 Statistical Features

Here, we describe our proposed features for selecting the correct candidate pair of source words for a blend. The features are intended to capture the properties of potential source words that make them good components of a blend, and thus are motivated by linguistic properties of blends and cognitive factors in their processing. In this work we are not trying to construct a cognitive model of human interpretation of blends. However, blends are coined by humans operating within certain cognitive conditions (such as constraints on memory and pattern recognition), and these restrictions may give rise to observable statistical properties in the resulting blends. Even though the task of identifying the source words of a blend is sometimes difficult for humans, a computational system should still be able to exploit these properties in automatically performing this task.

2.2.1 Recognizability (REC)

We propose a set of 5 features that relate to the ability of a language user to recognize the source words in a blend. First, it has been noted that frequent items tend to undergo subtractive word formation processes; specifically for blends, this is likely because frequent source words are more easily recognizable (Lehrer, 2003). To reflect this, we include $freq(w_1)$ and $freq(w_2)$, the frequency of each of the two candidate source words, as our first two REC features.

Recognizability of a source word also depends on how much of the word is present in the blend. This is not so much a matter of absolute amount, but rather an effect of neighborhood size. A source word's neighborhood consists of those words that share the portion of the source word that is contributed to the blend (Lehrer, 2003). A source word with a smaller neighborhood is more easily recognized because there are fewer potential words compatible with the blend. Gries (2006) bases his corpus-based study of blends on the related notion of *recognition point*—the minimum amount of material a reader or listener requires to uniquely identify a word. He approximates the recognition point of a word as the minimum length prefix/suffix such that the word is the most frequent in the set of all words which begin/end with

[3] Here we are using the terms *prefix* and *suffix* in the string sense as opposed to the affix morphology sense.

that prefix/suffix. This notion also clearly emphasizes a comparison between the candidate source word and other potential source words.

Following these ideas, we assume that a source word's recognizability is correlated with the degree to which its frequency dominates those of other words with the same prefix/suffix which the source word contributes to the blend We capture this with the following two REC features:

$$\frac{freq(w_1)}{freq(prefix)} \tag{1}$$

$$\frac{freq(w_2)}{freq(suffix)} \tag{2}$$

where $freq(prefix)$ [$freq(suffix)$] is the sum of the frequency of all words beginning with *prefix* [ending with *suffix*].

Recognizability may also be reflected in the degree to which the source words are associated with each other. Blends often describe a conjunction of concepts (e.g., a *spork* is a *spoon* and *fork*), or correspond to a phrasal sequence (e.g., *permalloy* is *permeable alloy*) (Algeo, 1977). We therefore expect that more recognizable blends will be those whose source words are used together relatively frequently. We include as our fifth REC feature the pointwise mutual information (PMI) of the source words, a measure that captures whether two words co-occur more often than expected by chance.[4] Here we assume that two words co-occur if they appear in the same sentence.

2.2.2 Contribution and Length (CONT/LEN)

We use 4 features which capture properties relating to the orthographic length of a blend's source words and the amount of material they contribute to the blend.

Our first CONT/LEN feature comes from Gries's (2004) finding that the second source word tends to be longer than the first, in terms of either graphemes or phonemes. We capture this as a feature according to the following formula where $len(w)$ is the length of w in graphemes.

$$\frac{len(w_2)}{len(w_1)+len(w_2)} \tag{3}$$

Gries also notes that the second source word will tend to contribute more of itself to the blend than the first source word. Considering the contribution in terms of graphemes, we first define the *contribution* of a word w to a blend b as:

$$cont(w,b) = \frac{len(prefix\ or\ suffix\ of\ w\ in\ b)}{len(w)} \tag{4}$$

We then define the second CONT/LEN feature as:

$$\frac{cont(w_2,b)}{cont(w_1,b)+cont(w_2,b)} \tag{5}$$

Gries further finds that the shorter source word will tend to contribute more than the longer source word. We encode this as the third CONT/LEN feature according to the formula below, which is positive when the shorter source word contributes more and negative otherwise.

$$[len(w_1)-len(w_2)]*[cont(w_2,b)-cont(w_1,b)] \tag{6}$$

Kubozono (1990) finds evidence that the length of a blend will tend to be similar to the length of its second source word. Our approximation yields the fourth CONT/LEN feature:

$$1-\frac{|len(w_2)-len(blend)|}{max(len(w_2),len(blend))} \tag{7}$$

2.2.3 Phonetics (PHON)

We include 3 features motivated by phonetic similarity. It is expected that source words are phonetically similar to the resulting blend (Gries, 2006). Since we would like our features to be calculable for truly novel blends for which we may not have a phonetic representation, here we approximate phonetic similarity by orthographic similarity. We calculate the longest common subsequence of each of the candidate source words and the blend for our first two PHON features.

The source words of a blend also often have a noticeable degree of phonetic overlap with each other (Algeo, 1977; Gries, 2006). We use the longest common subsequence of the two candidate source words as our third PHON feature. Here, as in Gries (2006), we use the phonetic transcription of the source words, since it is assumed that the source words (unlike the blend) are known words that do occur in a lexical resource.

2.2.4 Part of Speech (POS)

Since many blends are similar to conjunctions, we expect that the source words which form a blend will often be of the same part of speech (POS). Our single POS feature is an estimate of

[4] $PMI(x,y) = \frac{log p(x,y)}{p(x)p(y)}$

the probability of each candidate source word pair, w_1, w_2, occurring as the same coarse-grained POS according to the formula below.

$$\sum_{pos_i \in \{N,V,A,O\}} \frac{freq(w_1,pos_i)}{\Sigma pos_{j \in \{N,V,A,O\}} freq(w_1,pos_j)} * \frac{freq(w_2,pos_i)}{\Sigma pos_{j \in \{N,V,A,O\}} freq(w_2,pos_j)} \tag{8}$$

where $freq(w, pos)$ is the frequency of word w occurring as POS pos, and N, V, A, and O are noun, verb, adjective, and other, respectively.

2.2.5 Semantics (SEM)

Lehrer (2003) notes that people can more easily identify the source words of a blend when there is a semantic relation between them. We formulate two SEM features to incorporate this type of knowledge into our identification method.

As noted, blends are often composed of two semantically similar words, reflecting a conjunction of their concepts. Thus, a *spoon* and *fork* are both eating utensils that combine to form a *spork*, while *melt* and *weld* both involve the application of heat, combining to form *meld*. Our first SEM feature captures semantic similarity using an ontological similarity measure, which is calculated over an ontology populated with word frequencies from a corpus.

The other common type of blend corresponds to the blending of sequential phrases, as noted earlier in *permalloy*. Here, the source words are not necessarily similar, but are typically semantically related since they can form a felicitous phrase. Our second SEM feature is a measure of semantic relatedness using distributional similarity between word co-occurrence vectors.

2.3 Scoring the Candidate Pairs

To determine the source words for a blend, we sum the features described above, or some subset of them, to yield a score for each candidate pair.[5] In our experiments, we calculate the score separately using each of the 5 feature groups—REC, CONT/LEN, PHON, POS, SEM—as well as some combinations of them. Candidate pairs are then ranked according to the score given by the features under consideration, and some number of the highest ranked pairs are returned by the system.

Note that a candidate set may contain the same candidate pair multiple times, corresponding to different ways of splitting the blend, as in *boatel* formed from *boa(t)+(ho)tel* or *boat()+(hot)el*. Although the source words are the same, their split into a prefix and suffix differs, affecting some of our feature calculations. For such candidate pairs, we consider only the maximum score attained across its possible prefix–suffix splits.

3 Materials and Methods

3.1 Experimental Items

For this initial study, we wanted to use true blends that have been accepted by speakers, with source words objectively determined by lexicographers. Our experimental items thus are existing words from a standard lexical resource, the Macquarie dictionary (MD, Delbridge, 1981).

Using a simple regular expression, we automatically extracted all entries from MD whose etymology included the word *blend*, or indicated that the word was formed by combining two words, at least one of which was clipped. (The latter condition excludes compounds such as *broadband*.) The result of this process was a list of potential blends which we manually filtered to ensure that the items used in this study were truly blends. We removed any items for which either of the source words was an affix (i.e., *un-*, *-tion*) or did not have an entry in MD. We also eliminated any item whose lexical entry indicated doubt that it is a blend (e.g., a question mark at the beginning of the etymology field). To simplify the problem in this initial study, we imposed the additional constraint that the source words must not be words-with-spaces (e.g., *abietic acid*), and that no additional graphemes may occur between the source words in the blend (e.g., *donkophant*, *donkey* + *elephant*). From the resulting list, we manually extracted a total of 192 two-word sequential blends along with their corresponding source words.

The 192 blends were divided into 3 frequency ranges according to their frequency in the British National Corpus (BNC, Burnard, 2000). Development (DEV) and test (TEST) sets were selected to have an equal proportion of blends from these 3 groups, and contain 95 and 97 items respectively.[6]

[5] To ensure that the features have the same range of values and can be meaningfully summed, we normalize the features in each candidate set and take an arctan transform.

[6] In the experiments reported here, we do not consider the effect of the differing frequency ranges.

3.2 Candidate Sets and Features

To create the candidate sets we used all words in MD as our list of potential source words. We used frequency information from the BNC to calculate the word and prefix/suffix frequencies for the REC features. The CONT/LEN features were calculated from the spelling of the words. Phonetic transcriptions of each candidate source word were taken from MD, which we use (without stress markers) to calculate the PHON features. We use the BNC to determine the frequency of a word occurring as one of the 4 coarse-grained parts of speech for the POS features. We compute the first SEM feature, semantic similarity, using Jiang and Conrath's (1997) measure from the WordNet::Similarity package (Pedersen et al., 2004). The second SEM feature, semantic relatedness, is calculated as the cosine between word co-occurrence vectors, generated using software provided by Mohammad and Hirst (2006).

3.3 Evaluation Metrics

We evaluate the accuracy of our system as the percentage of items for which the correct candidate source word pair—as determined by the gold standard (MD)—is scored highest. However, this is a very stringent evaluation measure given the size of the candidate sets (ranging from 0 to over 2M, with averages on the datasets of 17K and 29K). To be of use in a semi-automated system, having the correct source word pair reasonably high in the scoring would be sufficient. We thus also evaluate our system according to an *in-top-5* accuracy metric, where the output of the system is deemed correct if the correct source word pair is in the top 5 pairs returned by the system. We compare these two system accuracies, *in-top-1* and *in-top-5*, to a uniform random baseline of $\frac{1}{\textit{size of candidate set}}$ and $\frac{5}{\textit{size of candidate set}}$, respectively.[7]

Given the importance of word frequency in blend formation, we also compare our system against an informed baseline which uses only two features—the frequency of each candidate source word. (Note that this baseline includes 2 of the 5 REC features.)

We do not have human judgments of source word identification on our blends, and so cannot provide an expected upper bound on performance for this particular set of blends. However, psycholinguistic research indicates that source word identification is a difficult task. Lehrer (2003) shows that human subjects achieve between 34% and 79% accuracy in source word identification, depending on the amount of material the source words contribute to the blend. When calculating relative error reduction, we consider the upper bound to be 100%, since the variation of human performance on this task does not indicate a clear upper bound.

4 Results on Source Word Identification

In Section 4.1 we report results on both DEV and TEST for identifying the source words from which a blend is formed. Because of the frequency of "conjunctive" blends (see Section 2.2), especially noun-noun blends like *brunch*, in Section 4.2 we also look at the performance of our methods on the subset of our blends formed from two nouns. We also have concerns that some of our experimental items are not actually blends.[8] In Section 4.3 we consider results on the subset of items which were confirmed to be blends in other lexical resources.

4.1 Results on All Blends

The first panel of Table 1 shows the results on DEV and TEST for both baselines and for the feature groups which outperform the frequency baseline, REC and SEM. The accuracies obtained using the CONT/LEN, PHON, and POS features are mostly better than the random baseline, but do not outperform the frequency baseline.

The first panel of Table 1 also shows the results on two combinations of features. We combine the REC features with the SEM features, since they perform best alone. This approach is also supported by the observation that individual source word recognizability and semantic compatibility work together in the human identification of blend source words (Lehrer, 2003). The results using REC+SEM are better than the frequency baseline, but are not substantially better than the REC features. We also combine all the features, shown as ALL. This gives results similar to those obtained for REC and REC+SEM.

In the rest of this section we focus on the REC, REC+SEM, and ALL features, since these give the best results. Indeed, given the low random baseline of 1% on the *in-top-1* task and a likely upper

[7] We eliminate duplicate word pairs from the candidate sets before calculating their size.

[8] For example, it does not match our intuition that *clash* is a blend of *clap* and *dash*.

bound on human performance of 34–79%, the accuracies of 12–26% using these feature groups are a promising start.

4.2 Results on Noun–noun Blends

Since some of our features, such as semantic similarity, are based on the notion that a blend's source words are likely to be of the same POS, we expected our methods would perform better on noun–noun blends. We consider a word to be a noun if its predominant POS in the BNC was noun. DEV$_{\text{N–N}}$ and TEST$_{\text{N–N}}$ include all of the blends formed from two nouns from DEV and TEST, respectively. Each new dataset contains 53 items.

Results on noun–noun blends are shown in the second panel of Table 1. The performance of the feature groups shown is better than the frequency baseline, and also always better than that of the corresponding experiment on DEV or TEST (first panel of Table 1). Note that the frequency baseline for noun–noun blends is generally higher than that for all blends. Nevertheless, when we consider relative error reduction over the frequency baseline, the performance of REC+SEM is always higher on noun–noun blends than on all blends. This indicates that this combination of features does in fact perform better on noun–noun blends.

4.3 Results on Confirmed Blends

We extracted all items from DEV and TEST that are noted as a blend in at least one of two other lexical resources (Soanes and Stevenson, 2004; Mish, 2003). This process resulted in 29 and 30 items in DEV$_{\text{Conf}}$ and TEST$_{\text{Conf}}$, respectively.

Results on these blends are shown in the third panel of Table 1. When comparing these results to those on all blends (first panel of Table 1), we see that for the random baseline, frequency baseline, and REC features, the results are quite close for the corresponding dataset (DEV or TEST) across the two panels. The very consistent values for these basic results across the full and reduced datasets indicate that although the Conf datasets are small they are reasonably representative.

The results using the combination of all features are not consistently higher across the DEV$_{\text{Conf}}$ and TEST$_{\text{Conf}}$ datasets. However, the results using the combination of REC+SEM stand out in that they show a substantial increase in performance for both datasets over results using just the REC features. There are two possible explanations for why this is the case. One possibility is that the items which were not confirmed to be blends in another resource are in fact not blends (i.e., the etymology entries in MD are incorrect). In this case, the better performance of the REC+SEM features would confirm that (true) blends are indeed formed from words which tend to be semantically similar. The second possibility is that the unconfirmed items actually are blends (i.e., the other lexical resources either did not include these items or did not correctly identify their etymology). In this case, the higher performance of the REC+SEM features would suggest that lexicographers are more likely to recognize that a word is a blend when its source words are semantically similar.

4.4 Discussion

We have shown that the REC features, especially in combination with the SEM features, achieve good results on all the blends, and even better results on subsets of blends with particular properties. The complementarity of REC and SEM is in line with observations of the importance of a semantic connection between source words when one of the words is difficult to recognize (Lehrer, 2003). Although our *in-top-1* numbers are low in absolute terms, they show reductions in error rate of up to 20%, on a task that is known to be very difficult for humans (with performance varying between 34% and 79% depending on the dataset). Our *in-top-5* performance, reaching near or over 50% in a number of cases, indicates promise for a semi-automated assistant for lexicographers.

It is perhaps not surprising that our rough phonetic approximations of PHON do not perform well in identifying source words; similarly, our POS features are grounded on the fact that many blends use source words that have the same POS, but many do not. It is more surprising that the CONT/LEN features do not perform well. Corpus studies (Gries, 2004, 2006; Kubozono, 1990) have found consistent patterns of source word contribution and length which our features are drawing on. Interestingly, Lehrer (2003) observes that the amount of contribution of a source word to a blend is only a factor in human identification of source words for *unknown* blends, not blends that are familiar to the subjects. It is not clear though why this would affect the resulting pattern of contribution across blends, which our score is based on.

Our performance may be limited because we

Table 1: % in-top-1 and in-top-5 accuracy on blends in DEV and TEST.

	All Blends				Noun–Noun Blends				Confirmed Blends			
	DEV		TEST		DEV$_{\text{N–N}}$		TEST$_{\text{N–N}}$		DEV$_{\text{Conf}}$		TEST$_{\text{Conf}}$	
Features	top1	top5	top1	top5	top1	top5	top1	top5	top1	top5	top1	top5
Rand baseline	1	6	1	6	2	7	1	7	2	8	1	5
Freq baseline	6	27	14	31	8	36	15	30	10	28	13	33
REC	12	34	18	43	15	43	19	49	17	38	17	43
SEM	12	25	15	33	-	-	-	-	-	-	-	-
REC+SEM	17	31	26	43	23	40	32	55	28	45	27	47
ALL	17	34	15	39	25	40	25	45	28	52	13	43

consider the blends without any contextual information. Since people can identify the source words of blends much more easily in context (Lehrer, 2003), we are likely missing important clues to the source words by taking this approach. A next step is to use the distributional similarity between the context of a blend and its source words (and their contexts) to address this issue.

5 Results for Blend Identification

Our focus here has been on identifying the source words of a word known to be a blend. However, given the connection between our features and the factors that influence the creation of blends (Section 2.2), these same features may be useful in determining whether a word is in fact a blend. Our features were designed to have higher values for the candidate pair which formed a blend, and lower values for others. We therefore expect that the top score for a candidate source word pair for a blend will be higher than the top score for a non-blend which has no correct source word pair.

To test this hypothesis we perform experiments in which we compare our candidate pair scores across both blends and non-blends. To calculate the scores here, we use the REC features only, since adding the semantic information did not help on all our blends. For the known blends we use our full DEV and TEST sets. For the non-blends, we created two datasets of non-blends, DEV$_{\text{NB}}$ and TEST$_{\text{NB}}$. These datasets were formed by randomly sampling words from MD which were not blends according to our heuristics. The words were chosen to yield the same proportion of items in the three frequency ranges (Section 3.1), and the same proportion of words from three length ranges, as those in DEV. The resulting DEV$_{\text{NB}}$ and TEST$_{\text{NB}}$ sets contain 98 and 99 words, respectively.

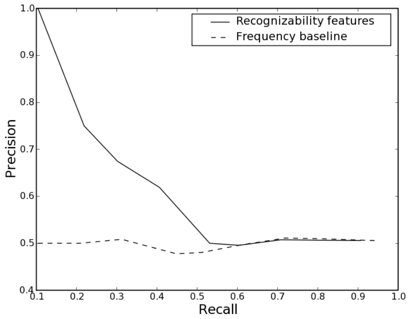

Figure 1: Interpolated precision-recall curve for REC features and frequency baseline on items in TEST and TEST$_{\text{NB}}$.

We created candidate sets and calculated the features for the items in DEV$_{\text{NB}}$ and TEST$_{\text{NB}}$ the same way we did for DEV and TEST. For each item, we then took the maximum of the score assigned to any of its candidate pairs.

The goal is to use our statistical score to separate blends from non-blends. We merge all items and determine how accurately the scores pick out the blends. That is, we rank all the blends and non-blends (together) by score, then consider higher scores to indicate that an item is a blend. We measure the precision for a fixed value of recall, and compare the results using the REC features against the frequency baseline features. The precision–recall curves for items in TEST and TEST$_{\text{NB}}$ are shown in Figure 1; results on DEV and DEV$_{\text{NB}}$ are similar. This figure shows that our method for determining whether an unknown word is a blend does tend to assign higher scores to blends than non-blends, and substantially outperforms the frequency baseline up through a recall value of 0.4.

6 Related Work

Much research in CL on unknown words has focused on determining various syntactic properties (e.g., Mikheev, 1997; Peng et al., 2004). Some work has tried to give a coarse semantic characterization of words unknown to the application (Toole, 2000). Our perspective here is quite different: we identify a highly productive word creation process (that of blending), and use the linguistic properties of blends to develop automatic means for their identification and interpretation. To our knowledge, we are the first to do so.

A related subtractive word formation process that has been investigated recently in the CL community is that of acronyming. Automatically inferring the expanded form of acronyms has received particular attention in the bio-medical domain. This problem is similar to that of identifying the source words of a blend in that the words in the expanded form must be identified from their orthographic contribution to the acronym. In fact, approaches to determining expanded forms, such as Schwartz and Hearst (2003) and Okazaki and Ananiadou (2006), are similar to ours in that they construct a set of candidate expanded forms and then attempt to select the most appropriate one from this set. However, since each word in an expanded form typically contributes only its first letter to the acronym, there is less information available to determine the expanded form than there is to infer the source words of a blend. On the other hand, the canonical form of acronym definitions and the domain specificity of the problem—two properties which have no counterpart in our work on lexical blends—can be exploited.

7 Conclusions

We identify a highly productive and very common novel word creation process—lexical blending—and investigate techniques for determining the source words of a blend based on the linguistic and cognitive properties of this process. Given a novel blend like *chillax*, the goal is to discover that this word is formed from *chill+relax*. This is a crucial step in determining the semantics of a large (and growing) class of neologisms. Our results indicate that statistical features capturing the recognizability of the source words consistently perform well in their identification. For blends with certain properties, performance is improved by the addition of features tapping into the semantic similarity and relatedness of the source words. Our results range from 12% to 32% on a task with an informed baseline of 6–15% and which human subjects find difficult. We also show that our features may be useful in determining that an unknown word is a blend.

Still, there is much room for improvement, as well as extensions to the work. First, we plan to explore the use of contextual information, which is critical to human performance on the source word identification task (Lehrer, 2003). Next, we must build on our work to tackle the next step of determining which sense each source word contributes to the blend, and how they are related, in order to achieve our goal of learning the semantics of a blend.

References

John Algeo. 1977. Blends, a structural and systemic view. *American Speech*, 52(1/2):47–64.

John Algeo, editor. 1991. *Fifty Years Among the New Words*. Cambridge University Press, Cambridge.

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

Arthur Delbridge, editor. 1981. *The Macquarie Dictionary*. Macquarie Library.

Stefan T. Gries. 2004. Shouldn't it be breakfunch? A quantitative analysis of the structure of blends. *Linguistics*, 42(3):639–67.

Stefan T. Gries. 2006. Cognitive determinants of subtractive word-formation processes: A corpus-based perspective. *Cognitive Linguistics*, 17(4):535–58.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*.

Anna Korhonen. 2002. Semantically motivated subcategorization acquisition. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pages 51–58.

Charles W. Kreidler. 1979. Creating new words by shortening. *English Linguistics*, 13:24–36.

Haruo Kubozono. 1990. Phonological constraints on blending in English as a case for phonologymorphology interface. *Yearbook of Morphology*, 3:1–20.

Adrienne Lehrer. 2003. Understanding trendy neologisms. *Italian Journal of Linguistics*, 15(2):369–382.

Diana McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations*. Ph.D. thesis, University of Sussex.

Andrei Mikheev. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

Frederick C. Mish, editor. 2003. *Merriam-Webster's Collegiate Dictionary*. Merriam-Webster, Incorporated, eleventh edition.

Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 35–43.

Naoaki Okazaki and Sophia Ananiadou. 2006. A term recognition approach to acronym recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (Coling-ACL 2006)*, pages 643–650.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *HLT-NAACL 2004: Demonstration Papers*, pages 38–41.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 562–568.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 474–479.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical texts. In *Proceedings of the Pacific Symposium on Biocomputing (PSB 2003)*, pages 451–462.

Catherine Soanes and Angus Stevenson, editors. 2004. *Concise Oxford English Dictionary*. Oxford University Press, eleventh edition.

Janine Toole. 2000. Categorizing unknown words: Using decision trees to identify names and misspellings. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 173–179.