A Proposal for the Automatic Distinction of Homomorphic Idiomatic and Non-idiomatic Phrases in WordNet

Benjamin R. Haskell
Cognitive Science Laboratory
Princeton University
ben@clarity.princeton.edu

Chandra Barnett
California Institute of Technology
chandra@caltech.edu

Christiane Fellbaum
Cognitive Science Laboratory
Princeton University
fellbaum@clarity.princeton.edu

Abstract

Idiomatic phrases composed of several lexemes pose various problems for NLP. It is often not obvious whether a given sequence of words is intended for idiomatic or literal interpretation. We propose a solution that detects idioms based on the semantic classes of their constituents. After annotating the idioms in WordNet with this information, they can be compiled into a tree structure to efficiently identify the constructions.

1 Introduction

Idiomatic phrases present multiple problems for NLP [Sag et al., 2002; Villavicencio et al., in press]. Structurally, they are composed of several lexemes, and a parser has to recognize them as a unit rather than a freely composed sequence of words. Once this is achieved, the idiom can be looked up in a lexical database such as WordNet and semantically interpreted. However, corpus data show that idioms occur far less frequently in a fixed syntactic configuration than is often assumed [Fellbaum and Stathi, 2006; Fellbaum and Geyken, 2005]. Moreover, modifiers such as adjectives and adverbs may be inserted into the idiom, making a simple string match impossible. The challenge is to recognize idioms in terms of their lexical components.

A hallmark of idioms is their lack of semantic compositionality, i.e., the meaning of the entire idiom is not the sum of its constituents [Nunberg et al., 1994; Jackendoff, 1995]. In fact, some idiom constituents, like *bygones* in *let bygones be bygones* and *gift horse* in *don't look a gift horse in the mouth* have no apparent independent meaning at all. These constituents never or rarely occur outside the idiom, thus posing no problems for semantic disambiguation. They can be spotted relatively easily by an automatic system and associated with the appropriate idiom syntactically and semantically in a fairly straightforward way.

On the next level of difficulty we find idioms with polysemous constituents that have at least one literal reading not associated with the idiom. Some idioms can be automatically identified with fairly high accuracy because they contain several lexemes that tend not to co-occur within a relatively small window outside of the idiomatic use. Examples are *kill*, *bird*, and *stone* (*kill two birds with one stone*) and *mountain* and *molehill* (*make a mountain out of a molehill*). In other cases, the idiomatic reading arises from a violation of the selectional restrictions associated with literal language: *lose one's heart/head, lose face*. Here, no literal reading is possible. These cases can nevertheless pose a challenge because the constituents are often very frequent and fairly polysemous; the mere co-occurrence of two lexemes would not suffice to identify the idiom. An automatic system would need to refer to a lexical resource that lists the selectional restrictions for the different senses of the verbs in some fashion.

Perhaps the most difficult cases are those where the string has both a literal and an idiomatic reading, as in *hit X on the nose*, *hit X on the head with Y, pour cold water on*, *let one's hair down*, and *have kittens*. We propose a solution for identifying the idiom in terms of the semantic category of the nouns in these idioms. The semantics will be specified as WordNet classes.

2 Scope

Some idioms are syntactically ill-formed and cannot be assigned to a syntactic category. An often-cited example is *by and large.* Many idioms are negative polarity items and require the presence of a negation: *not give a damn/hoot/dime, no use crying over spilled milk, no great shakes.* In this paper, we limit ourselves to idioms that fall into the conventional syntactic categories Noun Phrase, Verb Phrase, and Adjective Phrase. We exclude syntactic chunks and phrases like *cat's got your tongue* and *when it rains it pours*. One reason for our limitation is that our proposal is to match idioms, once identified, against WordNet, and WordNet recognizes only NPs, VPs, and APs. The work proposed here has not yet been carried out; we lack precise numbers but it appears that most idiomatic expressions fall within one of the three categories we consider.

We necessarily disregard constructions like *the Xer the Yer* and *what is X doing Y?* [Fillmore et al., 1988; Kay and Fillmore, 1999], which do not follow the syntax of the free language.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 177–180.

3 Idioms in WordNet

WordNet currently treats idioms as lexical units no differently from simplex words [Fellbaum, 1998]. Verb phrase and noun phrase idioms are entered as lexical units and linked to synonyms, hypernyms, etc. (these are usually simplex lexemes). But this kind of lexical entry pretends that idioms are fixed structures and do not occur in different configurations or with internal modification. One characteristic feature of idioms is that they may have open slots internal to the idiom. Typical is the the possessive, as in *keep one's ear to the ground*. Though we may include *keep one's ear to the ground* as a lexical unit, a token such as *Jeff always keeps* his *ears to the ground* will be difficult to process, as it does not contain the string as it is represented in the WordNet database, making a straightforward matching process impossible. However, if we instead enter the idiom as *keep [] ear to the ground*, and specify that the open slot must be filled by the possessive form of a personal pronoun, we may readily attach *keeps his ear to the ground* to the correct concept. (Indeed, a great number of idioms with the possessive genitive involve body parts or other inalienable possessions, and only pronouns coreferent with the subject or object occur in the determiner slot.)

In cases like the above example, matching text tokens against WordNet's lexical database could be practically accomplished by including the phrase with each of the possible pronouns as alternate forms in WordNet; the list is short and finite. However, in many common phrases, the open slots within an idiom may be occupied by any of a broad category of words. Fortunately, WordNet's built-in semantic relations provide us with a reasonable way to determine membership in these categories. If we envision the hyponymous relations between synsets in WordNet as forming a tree, we may specify an entire semantic class as the subtree rooted at some representative synset. Membership of a given synset in the semantic class can be efficiently determined by recursively checking direct hypernyms of the synset until we find either the representative synset or the root of the tree. Using this type of class abstraction, we may specify each open slot in an idiom as either requiring or excluding an entire semantic class. In this way, we can often determine whether an ambiguous sequence is intended in an idiomatic or literal sense. *John poured cold water on the team's plans* is probably intended for idiomatic reading, whereas *John poured cold water on the wilted tulips* is not. We may thus generate a decision tree that branches after *poured cold water on*. If the next token is found to be in the subtree of *{ object, physical object }*, we may conclude that a literal reading was intended; if it is in the subtree of *{ abstract entity }*, the idiom is implied.

For each idiom (or idiom class), the appropriate synset must be determined that subsumes all synsets whose members are candidates for the open slot. It is not desirable to include nodes above this synset. Thus, *{ abstract entity }* is probably not the best synset to characterize the slot in the idiomatic reading of *pour cold water on*. Corpus searches will reveal whether the tree should be cut at a node like *{ content, cognitive content, mental object }*, which subsumes concepts like *idea*, *plan*, and *design*.

For general applications, it will be helpful to pre-parse the input string. We will assume an intelligent parser that can resolve some common but non-canonical forms and generate a clause structure with the main verb and all of its arguments and adjuncts (some of which may be null), and any modifying clauses. In this way, we may branch first on the main verb, which offers a substantial early reduction to the size of the decision tree, instead of forcing us to analyze the terms in the order in which they occur in the input string. For example, differentiating *The cat had kittens* and *The boss had kittens* without pre-parsing is an awkward proposition — we must consider every idiom which specifies some superordinate of *cat* in the first position. After searching them all, and finding that *have kittens* is among them, we must then re-examine the subject to determine that subordinates of *{ cat, feline }* are excluded from this position, and *The cat had kittens* is in fact intended for a literal reading. If, however, we can first restrict ourselves to those idioms for which *have* is the main verb, we can more quickly identify *have kittens* as a possible idiom, which excludes subordinates of *{ cat, feline }* as the subject. By then determining whether or not the subject noun is found in a synset subordinate to *{ cat, feline }*, we will discover whether we are dealing with an agitated person or a lot of cats.

For a more complex example, suppose that the main verb is *hit*, which is a component of the idiomatic phrases *The film hit Joe over the head with the message* and *Lisa hit the idea on the nose*, as well as the similar but literally-intended *Lisa hit Joe over the head with a frying pan*. The decision tree might look like Figure 1 on page 179.

In the example *Lisa hit Joe on the head with a frying pan* the input to the decision tree would be something similar to Figure 2.

$$\left\{\begin{array}{ll} main\ verb & hit \\ subject & Lisa(person) \\ object & Joe(person) \\ location & head \\ instrument & frying\ pan \end{array}\right\}$$

Figure 2: Input for *Lisa hit Joe on the head with a frying pan*

Using the WordNet hierarchy, *frying pan* can be identified as a subordinate of *{ object, physical object }*. Thus, the process would follow the transitions *{ hit → person → on the head → with an object }* and conclude that the expression is not intended in the idiomatic sense.

Ultimately, all idiomatic expressions that pose disambiguation challenges would be compiled into a single decision tree, allowing fast lookups to be performed. Generally, the states leading to the determination of a *literal* reading can be omitted from the tree; failure to transition to a new state would indicate that a non-idiomatic sense had been encountered.

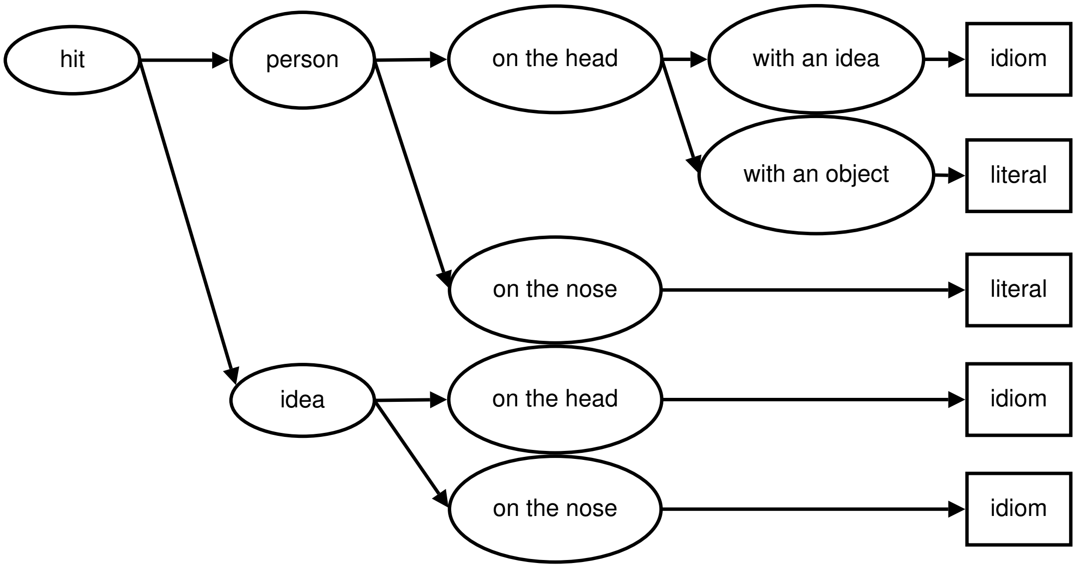

Figure 1: Decision tree for *A hit B on the C with D*

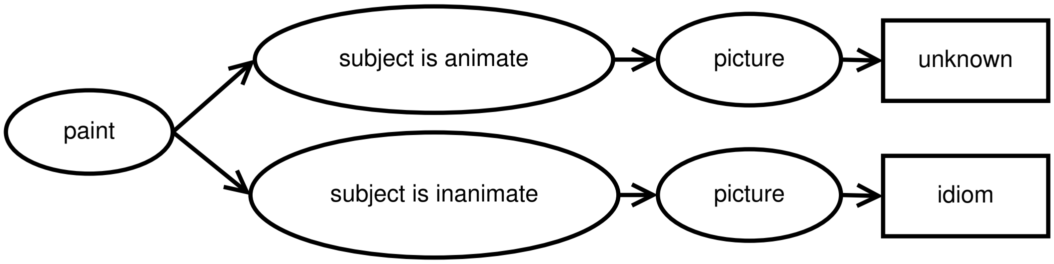

Figure 3: Decision tree for *X paints a picture of Y*

4 Limitations and open problems

An interesting problem arises for idioms like *paint a picture*, where some cases are clearly idiomatic, but others are ambiguous. If the subject is inanimate, as in *The brochure painted a rosy picture of the retirement home*, the idiom is intended, but in *He painted a pretty picture of the school*, it is unclear which reading should be chosen. This results in the decision tree in Figure 3.

There are other idioms for which our approach is completely unable to distinguish multiple readings. In *give X the axe*, no semantic class is sufficient to resolve the ambiguity between physically transferring an axe and firing an employee, and the larger context will have to take care of the disambiguation.

Another problem arises from the use of WordNet as the source of our semantic classes: many of the classes necessary for the slots in the idioms are difficult to define. Above, we specify the class of *{ object, physical object }* for the instrument that the hitter uses in order to distinguish the literal, compositional reading from the idiomatic reading, a better way to characterize this noun class would be in terms of a feature *[+solid]*, that is: something with the property of being solid. This feature covers entities in many WordNet classes, including *{ natural object }* and *{ artifact }*, that could occur in this slot, such as *rock*, *branch*, *book*, and *frying pan*. These same broad classes, however, also contain many nouns that cannot be the instrument of a hitting action, such as *constellations* and *air conditioning*. Extra corpus data might reveal better-defined classes to use in the decision trees for various idioms, and could in turn provide perspective on potential problems with the WordNet hierarchy itself.

Moreover, a good understanding of the semantic category of the nouns in idioms would provide insight into the constraints on lexical variations, where speakers substitute a context-specific for an idiom component [Fellbaum and Stathi, 2006].

Finally, one may question the use of WordNet as the lexical resource against which the strings are matched. For better or for worse, WordNet assumes an enumerative and finite sense inventory. However, viable lexica for NLP are bound to this undoubtedly oversimplified view of lexical semantics.

5 Summary and Conclusions

Coupled with a good parser, compiling idioms into decision trees with selectional slots based on WordNet's semantic classes could significantly improve the recognition of idioms. The success rate of this method can further provide a measure of the accuracy of WordNet's synset relations, and allow the WordNet hierarchy to be improved.

Acknowledgments

Barnett's work was made possible through a SURF grant from the California Institute of Technology and was carried out while the author was visiting the Cognitive Science Laboratory at Princeton University. Haskell and Fellbaum's work has been made possible by a grant from ARDA-AQUAINT. We thank the GWA reviewers for their helpful comments.

References

Christiane Fellbaum and Alexander Geyken. 2005. Transforming a corpus into a lexical resource for idioms and collocations. *Revue Française de Linguistique Appliquée*, X(2).

Christiane Fellbaum and Ekatherini Stathi. 2006. Ididome in der grammatik und im kontext: We brüllt hier die leviten? In Kristel Proost and Edeltraud Winkler, editors, *Festschrift für Gisela Harras*. de Gruyter, Berlin.

Christiane Fellbaum. 1998. Towards a representation of idioms in WordNet. In Sanda Harabagiu, editor, *COLING/ACL 1998 Workshop: Usage of WordNet in Natural Language Processing Systems*, pages 52–57, Montreal, Canada.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538, September.

Ray Jackendoff. 1995. The boundaries of the lexicon. In Martin Everaert, Erik-Jan van der Linder, André Schenk, and Rob Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, chapter 7. Lawrence Erlbaum Associates.

Paul Kay and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The what's x doing y? construction. *Language*, 75(1), March.

Geoffrey Nunberg, Ivan A. Sag, and Tom Wasow. 1994. Idioms. *Language*, 70(3):491–538, September.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, pages 1–15, Heidelberg/Berlin. Springer-Verlag.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. in press. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech and Language*.