Abstract

Sometimes a phrase, consisting of two or more words, stands out in the lexicon. A specific phrase may become the predominant way of saying something, such as with “strong tea”. One doesn’t hear tea described as “robust” or “stout”, although those adjectives could be used to describe tea that has more than the usual amount of flavor. Sometimes, a phrase may take on a meaning completely unlike any of the words that comprise it. How could someone not familiar with the phrase “kicked the bucket” know it means “died”? Phrases such as these are called Multi Word Expressions (MWEs). Many, if not all, human (natural) languages have them. MWEs cause problems for computers trying to generate natural-sounding language. They also cause problems for computers trying to parse and understand natural languages. Humans are quite adept at using MWEs, so computers performing natural language functions should be too. This paper describes the work done to improve computers’ abilities to handle MWEs.

1. Introduction

MWE have received much attention in scholarly literature for many years. I have discovered articles mentioning them all the way back to 1957 (Firth, 1957), and I’m sure they go back much further. One challenge for writing a review of this literature is not in finding relevant literature but in choosing which from among myriad articles to include. I’ll start by reporting how several clearly-written articles define MWEs (no one article says it all). Then I’ll cover articles that describe the problems that MWEs cause for computers trying to understand natural language. Next I’ll cover articles that describe various ways to mine corpora looking for MWEs. Then I’ll cover several landmark articles plus several good places to find other articles on MWEs. Lastly, I’ll mention some miscellaneous wisdom and thoughts on MWEs (mostly from the surveyed articles, but a bit of my own too.)

2. What is an MWE?

MWEs also appear to be known as idioms and collocations. In a graph intended to show the rise in importance of MWE research, Carlos Ramisch et al (Ramisch, Villavicencio, & Kordoni, 2013) have a figure (figure 1, p3:3) where they show the proportion of ACL Anthology articles from 1965 through 2006 containing the words “multiword”, “collocation” a/o “idiom”. The sum proportion rises (somewhat erratically) from 1965 through 2006. It is interesting to note that “idiom” is the most popular of those words until 1992, when “collocation” surpasses it. “Multiword” is fairly rare until 1999, when it surpasses “idiom” (which started a downward trend after “collocation” surpassed it.) Also in 1999, the sum proportion of those words starts increasing rapidly. In 2006, “multiword” had almost caught “collocation”. The first article I read about MWEs was actually a chapter in a book (Manning & Schutze, 1999). There, they called them collocations, and made no mention of “MWE” or “idiom”. The Ramisch et al figure may explain why Manning & Schutze called them what they did in the 1999 book.
Manning & Schutze mention noun phrase collocations such as “strong tea” and “weapons of mass destruction”. They also mention phrasal verb collocations such as “to make up”. They rhetorically ask why we say “stiff breeze” but not “stiff wind”, whereas we say both “strong breeze” and “strong wind”. They explain this as “convention” in the English language. (Other, more recent articles, give this type of collocation a specific name; more on that shortly.)

I’m confused when Manning & Schutze say that collocations have limited compositionality (where not-limited compositionality says the meaning of a phrase can be entirely determined by the meaning of the individual words.) I would say that anyone who knows the meaning(s) of “strong” and “tea” would immediately know what “strong tea” means. This, plus mentioning only a limited number of collocation types, is the only weakness I see in their treatment of collocations. They have an excellent introduction of how to find collocations in sample text; I will talk more about this in the appropriate section.

Of any MWE-article I read, the one that contained the most thorough yet succinct description of the various types of MWEs is “Multiword Expressions: A Pain in the Neck for NLP” (Sag, Baldwin, & Bond, 2002). If you only read one paper on MWE, this should be it! Not surprisingly, this was the most-often-cited paper of any I read on MWE. Besides referencing it frequently in this section, I will cover it under “Problems caused by MWE”.

Sag et al break MWEs into two broad categories: lexicalized phrases and institutionalized phrases. They further break the lexicalized phrases into three categories (or types of expressions): fixed, semi-fixed and syntactically-flexible. Semi-fixed expressions are further subdivided into non-decomposable idioms, compound nominals and proper names. Syntactically-flexible expressions are further subdivided into verb–particle constructions, decomposable idioms and light verbs.

Several examples of fixed expressions are: “by and large”, “ad hoc” and “in short”. Grammar rules don’t apply to fixed expressions; you can’t modify them to say “in shorter”, “in very short” or “by and small”.

Several examples of semi-fixed expressions are: “kick the bucket” (non-decomposable idiom), “part of speech” (compound nominal), and “San Francisco 49ers” (proper name). You can say “kicked the bucket” and “parts of speech” but not “kicked the buckets” nor “part of speeches”.

Several examples of syntactically-flexible expressions are: “brush up on” (verb–particle construction), “spill the beans” (decomposable idiom) and “make a mistake”/”do a demo” (light verbs). Permissible changes include past-tense versions: “Brushed up on”, “spilled the beans” and “made a mistake”, but not mixtures like “do a mistake” nor “make a demo”.

The institutionalized category is where Sag et al put “strong tea” and “telephone booth” (this is the category that Manning & Schutze just called “by convention”).

Sag et al actually made distinctions between MWEs, idioms and collocations; something that no other article did.

The last article I report on in this section is actually a slide set of a course given by one of the Sag et al authors: (Baldwin, 2004), and a prolific author of MWE articles himself. The 300+ slides in this set give a thorough refresher of many topics (including MWE examples on slides 4-20) covered in Sag et al, plus they tell the MWE story like only someone who’s been in the field for a long time can. Baldwin did a great job designing these slides so that a lot of information gets clearly communicated in a small amount of space. This is another must-read article on MWEs. It will make you wish you had attended the course in person.

3. Problems caused by MWE

Before saying what the problems are, I’d like to emphasize an often-cited source that implies how important it is to solve these problems. In his book ‘The Architecture of the Language Faculty” (Jackendoff, 1997, pp 154-158), Jackendoff considers the corpus of words/phrases used in the popular TV game show Wheel of Fortune. He calculates the number of those words/phrases used over the prior 10 years (all could be considered MWE), assumes that most Americans would be familiar with them, and estimates that the number of MWE that people know to be on the order of the number of single words people know. If this is the case, then for computers to be able to handle more than 50% of the language (i.e. words) that humans can, they need to know how to handle MWEs.
Ignoring MWEs would cap a computer’s NLP performance at a paltry 50%. In “Pain in the Neck…” (Sag et al., 2002 p2), they mention Jackendoff’s MWE estimate, then say it is likely higher. They reason that, beyond language in the general domain (i.e. what is found in Wheel of Fortune), there are many specialized domains (e.g. Bio, business, computer…) that add a greater proportion of MWE over simplex words (in these cases, institutionalized phrases) into what NLP applications should be aware of. Therefore, Sag et al propose that Jackendoff probably underestimates the number of MWE that people know.

### 3.1 Awkward sounding phrases

In Natural Language Generation (NLG), computers generate phrases for people to read. Those phrases often consist of adjectives that modify nouns. While an adjective may have many synonyms, that doesn’t mean that any of those synonyms would sound as natural while modifying a given noun. Take for instance Manning & Schütze’s discussion of tea. Powerful is a synonym of strong. If you said “powerful” tea, most English speakers would know what you mean, but saying “strong” tea sounds much more natural. So NLG software needs to be aware of MWE conventions to generate natural-sounding language.

### 3.2 Bloopers

A person commits a blooper when they say something that comically means something different than what they intended. One way to commit a blooper is to use an MWE in a way which is not conventionally accepted. Kenneth Church, in (Church, 2013 p4:3), relates an event where Winston Churchill used the MWE “put up with” in a non-accepted way, hence saying something that didn’t make sense and showing his ignorance in this matter. That MWE is a syntactically-flexible expression, where tense changes are allowed, but rearranging the words and placing other words in between them isn’t allowed. Yet Churchill rearranged (i.e. “…up with which I will not put”), and ended up sounding foolish. If computers try to generate natural language but don’t adhere to the rules of limited “flexing” of MWEs, they too will have the problem of looking foolish.

### 3.3 Words-with-spaces versus General compositional methods

Sag et al recognize that NLP can be done two different ways: using either statistical or symbolic methods. They give the edge to symbolic models, but say that symbolic models must solve two key problems if it is to become “linguistically precise”. One of those problems is handling MWE. (The other is performing disambiguation, but this doesn’t concern MWE.) The Sag article was published in 2002, so you may wonder if that problem has since been solved. From reading the articles in the 2013 issue of ACM Transactions on Speech and Language processing though, I’d say not. It appears that at least several promising techniques are still using statistical approaches, which leaves room for mistakes (at least until probabilities get extremely close to 100%). But symbolic techniques have their problems too. Sag et al talk about two ways that symbolic NLP software can treat MWE: either as words-with-spaces or using compositional methods. Depending on the type of MWE, one of the ways is better, and the other way may be very bad.

The words-with-spaces method treats an MWE as all one word. (In that respect, I believe a better method name would be “word-with-spaces”, since the method is treating the MWE as just one word.) If the MWE is a fixed expression (e.g. by and large or ad hoc), then words-with-spaces does well. But when the MWE can be constructed with a pattern (such as: take a walk, take a hike, take a flight), words-with-spaces fails to generalize well, and fails to predict what else could be an MWE.

The compositional method basically ignores the fact than an MWE is a special series of words. It would have a part of speech and a definition for “hoc” in “ad hoc”, even though it doesn’t make sense without “ad”. This results in what Sag et al call over-generation. Over-generation means that phrases that don’t make sense could be generated, like “telephone cabinet” to refer to a “telephone booth” (American) or “telephone box” (British/Australian.) And the compositional method can’t begin to handle idiomatic phrases like “kick the bucket”.

Sag et al propose various analytic techniques to mitigate the above-mentioned problems. These techniques are mostly compositional, but they do propose using statistical
techniques to handle the class of Institutional MWE.

3.2 Machine Translation

Ramisch et al provide a good example of why machine translation software needs to have semantic understanding of MWE in order to provide accurate translation. If software is trying to translate a sentence containing “Big Apple” into another language, it must know that “Big Apple” is an MWE meaning New York City.

3.2 Sentiment Analysis

Ramisch et al mention that it is valuable for software to be able to understand the semantics of MWE in the field of sentiment analysis. One of their examples presents the MWE of “having a riot”. While “riot” by itself has negative sentiment, the MWE actually means “having a good time” and hence has positive sentiment. I present this as another example of why software needs a semantic understanding of MWE. Suppose a computer is reading blogs that people write about the new iPhone, and wants to determine if people like it or not. One sentence contains the phrase “…the iPhone is a great deal,…”. In isolation, it may appear that the blogger likes the new iPhone. If we look further, the entire first part of the sentence may say: “If you think the iPhone is a great deal,…”. Ok, so now the computer knows the blog sentiment isn’t yet for or against the iPhone. The entire sentence is: “If you think the iPhone is a great deal, I have a bridge to sell you.” What is the computer to think? “I have a bridge to sell you” is a MWE, and if the computer knows what it means, it can correctly assign for-or-against to the blog.

There is a good article in the MWE-issue of Transactions of Speech and Language Processing (Klebanov, Burstein, & Madnani, 2013) on gauging sentiment of noun-noun pairs, even without trying to learn the meaning of the surrounding sentence. They talk about how to gauge sentiment of noun-noun pairs like heart attack. Their algorithm shows a good ability to gauge sentiment even when the meaning of the word-pair is noncompositional.

4. Mining Corpora for MWE

Manning and Schutze review several ways (created by others) to mine corpora for MWE. Their simplest technique, just looking for the most-frequent word pairings, fails (and they say so) because it is biased toward the most frequent single words (whether they are part of MWE or not). Then they talk about adding a POS filter that only accept phrases where words have specific parts of speech, like AN, NN or AAN (where A means adjective and N means noun. The MWE White House is an “AN” phrase.) These patterns were used by (Justeson & Katz, 1995), and work surprisingly well to identify likely collocations. The POS filter (plus broad enough corpus) is all that’s needed to show that strong tea is preferred over powerful tea. Interestingly, they say that most of the instances of powerful tea were in computational linguistics literature on collocations!

The following mining techniques discussed by Manning and Schutze eliminate potential MWE whose words aren’t most-frequently tagged as a verb, noun or adjective (so anything with “the” in it is eliminated). These techniques use statistical tests and can be used to discover MWE whose words aren’t necessarily contiguous in the phrase, like “knock … door”. (The question here is: should we say “knock on the door” or “hit”, “beat” or “rap” “on the door”, “at the door”, “on John’s door”,…?) To decide if we should use “knock”, “rap”, “hit”, or “beat”, we first scan the corpus looking for how many word positions, up to 4 on either side, “knock” occurs in relation to “door”. We do the same thing for “hit”, “beat” and “rap”. Calculate the mean and variance of these position-differences. A low variance indicates a possible MWE. They attribute this technique to Frank Smadja (Smadja, 1993).

Manning and Schutze add hypothesis testing to eliminate potential MWE where constituent words occur frequently in the corpus. They want to know if the words occur together more often than they should, given how frequently they may co-occur due to pure chance. They use the t test and Pearson’s chi-square test, among others, to do this.

Manning and Schutze mention the possibility of using mutual information to find MWE, but dismiss it as having limited utility.
One paper in particular in the MWE-extraction domain seems to stand out: (Smadja, 1993). The variance technique described by Manning and Schutze (above) is a simplified version of Frank Smadja’s work. Smadja uses a three-stage process, where the first stage is the statistical word association described above. The output of this stage feeds both stage two and three. The second stage identifies multiple-word combinations and complex expressions, with a claimed accuracy of ~40%. The third stage combines parsing and statistical techniques to label and filter possible collocations from stage one. The output of stage three is claimed to be ~80% accurate and have a recall of 94%. Smadja did this work with a 10-million word corpus of stock market news reports. His techniques have resulted in a lexicographic tool called Xtract. Smadja suggests that this tool has many applications, including language generation, retrieving grammatical collocations from corpora, deciding what type of determiner a noun should be used with, and machine translation.

5. Where to find good MWE articles

As I mentioned previously, Manning & Schutze contained the first MWE article I read. It gave a decent definition for MWEs (“collocations” was their word), then described some methods of searching corpora for MWEs in enough detail to make me want to start coding.

A search on Web of Science or maybe Google Scholar turned up two issues (2 & 3 of volume 10 in 2013) of ACM Transactions on Speech and Language Processing, devoted entirely to MWEs, starting with (Ramisch et al., 2013) and included 8 more articles.

There have been two other leading journals that have had special issues on MWEs: (Villavicencio, Bond, Korhonen, & McCarthy, 2005) and (Rayson, Piao, Sharoff, Evert, & Moiron, 2010).

The “Historical Overview” section of Ramisch et al pointed to several outstanding landmark articles for MWE. One was (CHOUEKA, 1988), which sadly, I was only able to find written in French. Another was (Smadja, 1993), also mentioned in Manning & Schutze as having a good way to extract collocations based on POS filters and word distance. Ramisch et al also pointed to another treasure trove of MWE articles: at the Stanford MWE project at http://mwe.sanford.edu. The Stanford trove contains pointers to the “pain in the neck” article by Sag et al, plus the Baldwin slide set and many others.

The “MWE Community” section of Ramisch et al also has pointers to good MWE sources. Manning & Schutze is mentioned there, as well as other NLP textbooks and MWE workshops and conferences.

Important MWE researchers include (but are not limited to): T. Baldwin, K. Church, R Jackendoff, S. Kim, C. Manning, K. McKeown, I. Sag and F. Smadja.

6. MWE Wisdom

Many articles gave advice on how to better parse text with MWEs in mind, or on data structures that could represent knowledge contained in MWEs. In this section though, I wish to report on things that humans’ use of MWE says about how our brains are wired, plus give a tip about a simple way to “say it right” in a language you may not be intimately familiar with.

First, in (Manning & Schutze, 1999, p152-3) say “It is a convention in English to talk about strong tea, not powerful tea, although any speaker of English would also understand the latter unconventional expression.” They go on to ask “why do we use powerful for drugs like heroin, but not for cigarettes, tea and coffee?” They posit that the distinction is important as a way to inform our society of important lessons (i.e. “strong” is OK but “powerful” must be used with caution if at all).

Along similar lines, I say that the existence of institutional sayings (accepted ways of saying things) shows we, as a race, are willing to conform to social norms.

When we create expressions like “He talked ad nauseam...” to shorten “He talked so long about... it made me sick”, we aim to get our thoughts communicated in as few words as possible.

When we replace “He died” with “He kicked the bucket” we try to lessen painful thoughts.

When we replace “She played the piano” with “She tickled the ivories”, we want to communicate our thoughts in unique and interesting ways.
Here’s a tip about how to always say a phrase the conventional way. Suppose you want to ask Starbucks for a cup of coffee that has a lot of flavor. You now know that flavorful tea is referred to as *strong* tea, but does the same convention hold for coffee in the English language? Should you tell them you want it “strong”, or maybe “robust”, “tough”, or “durable” or with some other synonym for *strong*? As it turns out, Google developed a tool that can help us with those kinds of questions (Michel et al., 2011). (Interestingly though, I have never seen any literature mention using this tool to help with MWE discovery.) Google digitized millions of books in many different languages, and provides a way to view n-grams from them all. Summaries of those n-grams show the “collective right way” to say most anything you can think of. Go to https://books.google.com/ngrams/ and type in your four different ways to say flavorful coffee (i.e. “strong coffee, robust coffee, tough coffee, durable coffee”). You will see the following chart:

![](image)

Since one way stands out above all the others, it tells you that “strong coffee” is the way the vast majority of people said it (in the books Google scanned), which indicates that “strong coffee” is the right MWE to use. This trick works for comparing any phrases that contain up to 5 words. It also works in many other languages (including French, German, Chinese...). To find the most common words that come before and after a given phrase, just put an asterisk (i.e. “*”) in those positions. Knowing whether to use “the” in front of a noun is difficult for many people who aren’t native English speakers. So if you want to know if “the” should be used before a given noun in some phrase, enter “* <noun> <phrases>*” into the search box, then observe if “the” is mentioned in the results.

References


