Word Senses

What is a Word Sense?

If you look up the meaning of word up in comprehensive reference, such as the Oxford English Dictionary (the OED), it is usually the case that the word has several senses, often spread across multiple parts of speech. For example, in the most recent edition of the OED, the word "run" has fifteen senses in adjective form, over fifty senses in noun form, and over eighty senses in verb form! The task of choosing which word sense most accurately represents the sense of a particular use of a word is known as *Word Sense Disambiguation* (or WSD). It is the process of matching up words in a text with their corresponding sense entries in the dictionary.

In the examples in this guide, and in most of our annotation projects, we will be drawing our potential senses from an electronic dictionary called *Wordnet* that was assembled by linguists at Princeton University over the past three decades. Wordnet is organized slightly differently than a regular dictionary: its main entries are called *synsets*, short for *synonym sets*, which is a single word meaning that might apply to a number of different words. For example, there is synset, identified by the number 06626039, which means "a form of entertainment that enacts a story by sound and a sequence of images giving the illusion of continuous movement." Its synonyms are "movie," "film," and "picture." On the other hand, there is another synset, 03882197, meaning "graphic art consisting of an artistic composition made by applying paints to a surface," whose synonyms are "painting" and "picture. Whenever we encounter the word "picture" we must determine if one or the other synset applies (or possibly a different one altogether).

Preliminaries

Before we can actually assign senses to words, we must determine, at the very least, where the words are in the text. This is not as easy as it might first appear. We must also determine the root form and the part of speech of the word, so we can effectively search the dictionary for appropriate senses.

Determining Word Extent

The main complication in determining the where words are is the presence of *multi-word expressions* (MWEs). A multi-word expression, also known as a *collocation*, is a group of tokens that (1) occurs together more often than would be expected by pure chance and, (2) is arbitrarily restricted with regard to its syntactic or semantic flexibility. MWEs make up anywhere from 10 to 30% of the words in a text, on average. Examples of common MWEs are compound nouns such as "world record," verb-particle constructions such as "look up", or proper nouns, such as "Guinness Book of World Records." For example the following sentence contains three MWEs:

(1) She looked up the world record in the Guinness Book of World Records.

Keep in mind the distinction between *tokens* and *words*. Tokens are continuous runs of non-whitespace characters, which may or may be separated from the neighboring tokens by whitespace. In the above example there are 14 tokens, which include every individual word plus the period at the end of the sentence. Certain sequences of characters that are usually thought of as a single word are actually split into two tokens for ease and uniformity of analysis. Contractions such as "can't" and "won't" almost always fall into this category: "won't" becomes two

tokens "wo" and "n't", analogous to the two tokens present in its uncontracted form, "will not." The "wo" is considered to be an inflected form of "will."

While the vast majority of MWEs are made up of contiguous sets of tokens, consider the following example:

(2) She looked$_1$ the word up$_1$.

Here, there is a single MWE, "looked up" that has two *interstitial* tokens, "the" and "word," which intervene between the MWE's first and second token. Regardless of this MWE being interrupted, it is still a valid multi-word expression.

How does one determine whether one or more tokens make up an MWE? There are three main criteria for determining whether a set of tokens is an MWE:

1. The set of tokens make up a proper noun, such as a person, place, or group.
2. The MWE is listed in the dictionary, in the sense as used in the particular context.
3. The set of tokens express a "non-compositional" meaning.

For the second criterion, contrast the next example with (1):

(3) She looked up the mountain.

While this sentence also contains the sequence of tokens "looked" followed by "up," it does not contain a MWE, because "looked up" isn't used in the same sense under which "look up" is listed in the dictionary, that of finding an entry in an index. Rather, it means physically looking in a particular direction. This can be seen by substituting "down" for "up" in both sentences. Example (3) remains intelligible, but (1) becomes nonsensical. This leads to a technique for determining whether the third criterion hold, that of non-compositionality. **Rule 1: When testing for non-compositionality, try replacing word in the MWE by similar tokens with identical parts of speech. If the phrase still makes sense, then it is probably not a MWE.** This rule works especially well with verb-particle constructions such as "look up" where the second word, the particle, can be easily replaced by other prepositions.

Marking the Part of Speech

Once we have found all the words, we must mark the parts of speech. Every token in the text should have a part of speech, plus every multi-word expression. Most people think of parts of speech as only have four categories: noun, verb, adjective, and adverb. But there are numerous parts of speech, and they vary between languages. Table 1 lists a common *tagset*, or collection of part of speech tags, used for English: it contains 48 entries. Twelve of the entries cover tokens that are punctuation, but the rest cover types of tokens we expect to find in dictionaries, such as noun and verb. You should study this table carefully and make sure you understand what all the tags mean.

There are three pairs of tags that can be difficult to distinguish. First, and most common, are two similar verb tags, VBD (past tense verb) and VBN (past participle). The past participle is used to form perfect constructions and speak in the passive voice, as in:

(4) *The chicken has eaten.* (perfect construction)
(5) The chicken was eaten. (passive voice)

Furthermore, it can be used to modify a verb or sentence, with passive sense

(6) Seen from this perspective, the problem presents no easy solution.

Tag	Description	Examples
CC	Coordinating conjuction	and, but
CD	Cardinal number	2, two
DT	Determiner	the, a, an
EX	Existential there	[There] is
FW	Foreign word	
IN	Preposition/subordinating conjunction	
JJ	Adjective	small, big
JJR	Comparative adjective	smaller, bigger
JJS	Superlative adjective	smallest, biggest
LS	List item	List item marker
MD	Modal	would, could, should, have
NN	Singular or mass noun	
NNS	Plural noun	
NNP	Singular proper noun	
NNPS	Plural proper noun	
PDT	Predeterminer	all, both, half, many, quite, such, sure, this
POS	Possessive ending	's
PRP	Personal pronoun,	he, she, it
PRP$	Possessive pronoun	his, hers, its
RB	Adverb	quickly, well
RBR	Comparative adverb	more quickly, better
RBS	Superlative adverb	most quickly, best
RP	Particle	move [up]
SYM	Symbol (mathematical or scientific)	+, -, >, <
TO	to-word	[to] eat, [to] love
UH	Interjection	amen, huh, howdy, dammit, shucks, anyways, honey, baby
VB	Base form verb	talk, hire, eat
VBD	Past tense verb	talked, hired, ate
VBG	Gerund/present participle verb	talking, hiring, eating
VBN	Past participle verb	talked, hired, eaten
VBP	Non-3rd person singular present verb	I use, you speak
VBZ	3rd person singular present verb	He uses, she speaks
WDT	Wh-determiner	that, what, whatever, which, whichever
WP	Wh-pronoun	that, what, whatever, which, whatsoever, who, whom, whosoever
WP$	Possessive wh-pronoun	whose
WRB	Wh-adverb	how, however, whence, whenever, where, whereby, wherever, wherein, whereof, why
#	Pound sign	
$	Dollar sign	
.	Sentence-final punctuation	[.], [!], [?]
,	Comma	
:	Colon or semi-colon	[:], [;]
(	Left bracket character	[, (, {, <
)	Right bracket character	],), }, >
"	Straight double quote	
`	Left single quote	
``	Left double quote	
'	Right single quote	
"	Right double quote	

Table 1: The 48 Part of Speech tags

Second, IN (preposition) and RP (particle) often look extremely similar in the case of phrasal verbs such as "look up." In this case, the "up" is a particle, not a preposition.

Third, very rarely, there is confusion between CD (cardinal number) and LS (list item). When you have an itemized list, such as "I like three types of pets, (1) cats, (2) dogs, and (3) fish," the numbers in parentheses are list items, but the "three" is a cardinal number.

Finding the Root Form

Once we have located all the words, both the single words and multi-words, and marked all their tokens (and the multi-word themselves) with their parts of speech, we must determine the *lemma*, or *root form*, of each word. The lemma is considered the "base" form of the word, with no inflection. For example, the words "ran," "running," and "runs" all have the same base form of "run." This is the form of the word under which one would expect to find its senses listed in a standard dictionary. The singular form of nouns is considered the base form of a word, e.g., the lemma of "whales" is "whale." The one exception to this rule is that Proper Nouns are their own root form.

Multi-word expressions also have lemmas, but one must often think carefully about what the base form is, as some multi-words inflect in unusual ways. This is often encountered in the plural form of multi-word nouns; for example, the plural form of "attorney general" is "attorneys general."

Keep in mind, also, that one cannot always determine the lemma without have a pretty good idea of what the sense of the word is. Take the surface form word "stove." This might be a noun meaning the kitchen appliance, in which case its lemma is the same as the surface form. On the other hand, if it is being used as a verb, its lemma is "stave," as in "I stave in his head."

Choosing a Word Sense

Once we have determined the extent of all the words, their lemmas and parts of speech, we are ready to assign a word sense to each word. In most cases, the appropriate sense will be apparent by carefully reading through the list of available senses. Each synset has as associated *gloss* and set of examples which illustrate the use of the word that can help you make a decision. Generally, the more common the word, the more senses it has, and initially you will have to spend a lot of time to read each sense carefully, to determine which is most appropriate in the given context. Over time, you will become familiar with the most commonly used senses and you will be able to assign them without requiring a lot of time.

While assigning senses is usually fairly straightforward, there are cases where the distinctions can be quite subtle. In these cases it is helpful to remember that the gloss is not a definition, it is merely a short description that makes it easier to distinguish that sense from other, similar senses for the synonyms in the set. It is merely suggestive. What is definitive is the set of actual synonyms: in almost every case, if you cannot plausibly replace a word with one of the synonyms in a synset, and retain the meaning of the text, then that synset is probably not the correct sense.

Consider the following example and two possible synsets, with their glosses and examples:

(7) The water poured all over the floor.

a. pour (pour%2:38:03::) cause to run; "pour water over the floor"

b. pour (pour%2:38:02::) flow in a spurt; "Water poured all over the floor."

Which synset should you choose for "poured" in the example? There are three guidelines that can be applied here to determine that (b) is the correct sense.

First, if we have an example in the gloss that matches the sentence *exactly*, that is usually the correct sense.

Second, senses often distinguish between active and passive forms of the same verb. In this case we have a passive form, rather than someone actively pouring the water, so the synset that emphasizes the passive is the correct on in this case.

Third, if all else fails, we can look at the id strings of the senses (the strings that have the percent signs in them). The senses are usually organized from more general to more specific, with more general senses assigned lower numbers. All things being equal, we should choose the more general sense.

There is a fourth guideline, one that relies on implicit and explicit information. Suppose two synsets are nearly indistinguishable, differing only by the presence of an additional constraint on one of the synsets. If the expression containing the word contains that additional constraint explicitly, then we should choose the synset without the constraint. This is illustrated as follows:

(8) Need a good example of this.
(9) Ditto.

Sometimes there will be no appropriate sense listed in the dictionary. When you encounter this, you should scour the dictionary to find another sense listed under a different lemma that would fit the context. It in extremely rare cases there will be appropriate synset *anywhere* in Wordnet. In these cases you are allowed to mark "no appropriate sense."

Finally, proper nouns are treated somewhat differently from words. For these, you are given the option of three generic senses, "person," "location," or "group." These will suffice for most proper nouns. Rarely, the proper noun will be listed in the dictionary (for example, many US presidents are listed in the dictionary), in which case you should pick that sense.

Summary of Rules

#	**Rule**
1	

Glossary

co-location – see *multi-word expression*

lemma – definition

multi-word expression (MWE) – definition

part of speech – definition

root form - see *lemma*

sense – definition

surface form – definition

synset – definition

token – definition

word-sense disambiguation (WSD) – definition

Wordnet – definition