Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone

Michael Lesk

Bell Communications Research
Morristown, NJ 07960

The meaning of an English word can vary widely depending on which sense is intended. Does a *fireman* feed fires or put them out? It depends on whether or not he is on a steam locomotive. I am trying to decide automatically which sense of a word is intended (in written English) by using machine readable dictionaries, and looking for words in the sense definitions that overlap words in the definition of nearby words.

The problem of deciding which sense of a word was intended by the writer is an important problem in information retrieval systems. At present most retrieval systems rely on manual indexing; if this is to be replaced with automatic text processing, it would be very desirable to recognize the correct sense of each word as often as possible. Previous work has generally either suggested (a) detailed frames describing the particular word senses,[1,2] or (b) global statistics about the word occurrences.[3] The first has not yet been made available in any real application, and the second may give the wrong answer in specific local instances. This procedure uses available dictionaries, so that it will process any text; and uses solely the immediate context.

To consider the example in the title, look at the definition of *pine* in the Oxford Advanced Learner's Dictionary of Current English: there are, of course, two major senses, "kind of evergreen tree with needle-shaped leaves..." and "waste away through sorrow or illness..." And *cone* has three separate definitions: "solid body which narrows to a point ...," "something of this shape whether solid or hollow...," and "fruit of certain evergreen trees..." Note that both *evergreen* and *tree* are common to two of the sense definitions: thus a program could guess that if the two words *pine cone* appear together, the likely senses are those of the tree and its fruit.

Here is the output:

```
pine  1*  7 kinds of evergreen tree with needle-shaped
             evergreen(1) tree(6)
      2   1 pine/ -
             pine(1)
      3   0  waste away through sorrow or illness:

      4   0  / pine for sth; pine to do sth, /  have a

cone  1   0  solid body which narrows to a point from a

      2   0  sth of this shape whether solid or hollow,

      3*  8  fruit of certain evergreen trees (fir, pine,
             evergreen(1) tree(6) pine(1)
```

Multiple word senses, whether semantic or syntactic, are a major problem in many areas of processing natural languages on computers. What researcher in NLP does not remember *time flies like an arrow?* Conventionally, with such examples parsers throw up their hands, and invoke hypothetical expert systems or complete models of the world. This paper is an attempt at a cheap solution to the problem of sense discrimination.

What we try is to guess the correct word sense by counting overlaps between dictionary definitions of the various senses. Look at the definitions of *ash* in Webster's 7th Collegiate:

```
ash
 1 any of a genus (Fraxinus) of trees of the olive family
   with pinnate leaves, thin furrowed bark, and gray
   branchlets
 2 the tough elastic wood of an ash
ash
 1 the solid residue left when combustible material is
   thoroughly burned or is oxidized by chemical means
   fine particles of mineral matter from a volcanic vent
 2 ruins
 3 the remains of the dead human body after cremation or
   disintegration
 4 something that symbolizes grief, repentance, or
   humiliation
 5 deathly pallor
ash
 0 to convert into ash
```

But now suppose the previous word is *coal*, which is defined as follows:

```
coal
 1 a piece of glowing carbon or charred wood : ember
 2 charcoal
 3 a black or brownish black solid combustible substance
   formed by the partial decomposition of vegetable
   matter without free access of air and under the
   influence of moisture and often increased pressure
   and temperature that is widely used as a natural fuel
   pieces or a quantity of the fuel broken up for burning
coal
 1 to burn to charcoal : char
 2 to supply with coal
 0 to take in coal
```

Leaving aside the difficulty of recognizing that sense (3) above is the common mineral coal on which the Industrial Revolution was based, note that this sense of *coal* uses in its definition the words *combustible*, *burn*, and *solid* all of which also appear in the definition of *ash* as "the solid residue left when combustible matter is thoroughly burned." So the program, by counting overlaps, can guess that *coal ash* involves this meaning of *ash*, rather than anything to do with the tree or the color. There is one overlap of the word *wood* between the first sense of *coal* ("a piece of glowing carbon or charred wood") and the second sense of *ash* ("tough elastic wood of an ash"), but only one; so the three overlaps win out. If one wished to describe the residue left when a fire of ash wood was burned, some circumlocution (such as just given) would be necessary.

Now look at the output of the program on the two examples *Time flies like an arrow* and *Fruit flies like a banana*. The columns give, respectively, the words in the sentences, the number of the sense, the count of the overlaps, and the first few characters of the sense definition. The matching words are shown below the sense definition. The starred sense number is one with the most matches, and therefore the selected one. The dictionary here is the Oxford Advanced Learner's Dictionary of Current English:

```
Word Sense Count      Definition & Matches
time    1    0  all the days of the past, present and
        2    2  the passing of all the days, months and
        3    6  (also / a - / - adj / + time/   portion
        4    2  point of time stated in hours and minutes
        5    2  time measured in units (years, months,
        6*   8  point or period of time associated with,
        7    0  (Cf twice/   occasion: / . this/that
        8    0  used to indicate multiplication (but
fly     1    2  two-winged insect, the common housefly
        2    0 fly/ -
        3    0  move through the air as a bird does, or
        4    1  direct or control the flight of (aircraft)
        5*  12  go or move quickly; rush along; pass
        6    0  cause (a kite/   to rise and stay high
        7    0  flee from: / . fly the
        8    0 fly/ -
        9    0  (also, colloq, / - pl /  used with /
       10    0  flap of canvas at the entrance to a tent
       11    1  (old use) one-horse hackney carriage.
       12    0  outer edge of a flag farthest from the
       13    0 fly/ - adj/  sl) /  cunning; alert; not
like           ---
an             ---
arrow   1    3  thin, pointed stick (...) shot from a bow
        2*   5  mark or sign ('/   used to show directio~
```

```
fruit   1*   9  (collective / - n/   that part of a plant
        2    0  (bot) that part of any plant in which
        3    1  / the fruits of the earth, /  those plant
        4    2  (fig, often -pl) profit, result or rewarc
        5    3  fruit-machine  GB colloq) coin-operated
        6    1  of or like fruit in smell or taste.
        7    0  (colloq) full of rough (often suggestive)
        8    0  (colloq) rich; mellow; florid: / a fruity
fly     1*   3  two-winged insect, the common housefly
        2    1 fly/ -
        3    2  move through the air as a bird does, or
        4    0  direct or control the flight of (aircraft)
        5    1  go or move quickly; rush along; pass
        6    0  cause (a kite/   to rise and stay high
        7    2  flee from: / . fly the
        8    1 fly/ -
        9    1  (also, colloq, / - pl /  used with /
       10    0  flap of canvas at the entrance to a tent
       11    0  (old use) one-horse hackney carriage.
       12    0  outer edge of a flag farthest from the
       13    1 fly/ - adj/  sl) /  cunning; alert; not
like           ---
a              ---
banana  1*   6 long, thick-skinned (yellow when ripe)
```

Note that the program has correctly distinguished *fly* in each case, although it has the wrong meaning of *arrow* (and an incorrect but arguable meaning of *time*). The reason it works is that *fruit fly* is actually in this dictionary buried under *fruit* to provide the necessary overlaps.

What are the advantages of this technique? It is non-syntactic, and thus a useful supplement to syntactically based resolution. For example, in *I know a hawk from a handsaw* this program is useless at telling *hawk* ("strong, swift, keen-sighted bird of prey") from *hawk* ("offer goods for sale"), but syntax would help immediately since the verb can not appear immediately after an article. But often syntax is of no help. Here are three different meanings of *mole* as a noun: *I have a mole on my skin; there is a mole tunnelling in my lawn;* and *they built a mole to stop the waves*. The program will do all of these.

Another major advantage is that it is not dependent on global information. Here is a sentence from *Moby Dick* (the unchosen sense definitions have been suppressed to save space):

```
There             ---
now               ---
is                ---
your              ---
insular    1*  4  of or like islanders; narrow-minded:
city       2*  9  (attrib): / city centre/ central area
of                ---
the               ---
Manhattoes        ???
belted     2* 86  any wide strip or band, surrounding
round      4* 28  (compounds) / round-arm / - adj, adv
by                ---
wharfs     0*  1 (or wharves) wooden or stone structure
as                ---
Indian     2* 16  (various uses) / Indian club, /
isles      0*  3 island (not much used in prose, except
by                ---
coral      0* 15  hard, red, pink or white substance
reefs      1*  9 ridge of rock, shingle etc just below
commerce   0*  1 trade (esp between countries); the
surrounds  0*  0 be, go, all round, shut in on all sides
it                ---
with              ---
her               ---
surf       0*  5 waves breaking in white foam on the
```

Note that it got the correct meaning of *reef*: the alternative meaning is *all hands to reef topsails*. If one depended on global information, one would conclude that since *reef* appears nine times in *Moby Dick* and seven of those are related to sails, that should be the meaning chosen; and the two instances of *coral reef* would be mistakes.

What properties should a machine-readable dictionary have to make this method work as well as possible? Experiments are being run on several, but I suspect the answer is simply bulk of information. Here is a comparison of some machine-readable dictionaries now available with the forthcoming OED:

	Size of Dictionaries			
	OALDCE	W7	CED	OED
Bytes	6.6 MB	15.6 MB	21.3 MB	350.0 MB
Headwords	21,000	69,000	85,000	304,000
Senses	36,000	140,000	159,000	587,000
Bytes/headword	290	226	251	1,200

OALDCE = Oxford Advanced Learner's Dictionary of Current English
W7 = Merriam-Webster 7th New Collegiate
CED = Collins English Dictionary
OED = Oxford English Dictionary (estimated)

To compare the dictionaries, consider *galley*. The second sense in the Oxford Advanced Learner's dictionary is:

2. ship's kitchen.

The entry in Webster's 7th Collegiate for sense 2 is somewhat longer, reading

4. the kitchen and cooking apparatus esp. of a ship or airplane

By comparison, the OED entry for *galley* for the same sense also includes words such as *stove, cook, cooking-room,* and *pot*. (The OED, of course, omits *airplane*). My program, running with the OALDCE and processing *stoke the stove in the galley*, chooses the wrong meaning of *galley*. *Stove* does not overlap with any meaning of *galley*; in the OED it would. So I eagerly await the OED, which with five times as much material per headword should do a better job at discriminating among the senses. In the meantime, experiments are being run to compare the other machine readable dictionaries and see how well they do on this algorithm.

What is the current performance of this program? Some very brief experimentation with my program has yielded accuracies of 50-70% on short samples of *Pride and Prejudice* and an Associated Press news story. Considerably more work is needed both to improve the program and to do more thorough evaluation. Meanwhile an attempt is being made to select good default options.

There are many interesting choices to be made. Which machine readable dictionary should be used? I have tried three. So far, it appears that results are roughly comparable with Webster's 7th Collegiate, the Collins English Dictionary, and the Oxford Advanced Learner's Dictionary of Current English. As explained above, I think that total length of entry will turn out to be the dominating characteristic. All of these are about the same.

Should compound words and their definitions be included and used? Sometimes they are relevant, and sometimes not. The definitions and uses of *money-lender* are obviously helpful in elucidating *money*; but the definitions and uses of *red herring* are not going to help with *red*. Can these be told apart automatically? Perhaps not, but then the Oxford Dictionary of Current Idiomatic English is available in machine readable form, so it should be possible to distinguish the true idioms.

Should examples be included and used? Examples can range rather widely afield from the root word. The OED's most recent example for *locust* is from Disraeli: *White ants may sink a fleet, or locusts erase a province.* And yet in my tests, the problems with both compounds and examples are not that they digress excessively, but that they are simply not much use because they add to definitions that were probably long enough already. Where the program needs help to amplify the definitions of words despatched by the lexicographer with a terse synonym, there are usually no examples or compounds listed.

Should the overlaps be counted and a numerical score used, or should one occurrence be as important as five? The program has this option, and so far it seems to make little difference. Should the overlaps be weighted by the length of the dictionary entry? Again, this option doesn't seem to mean much in practice. Considerably more evaluation is going to be needed to decide on the value of many of these options.

How wide a span of words should be counted? The program uses ten words as its default window; changing this to 4, 6 or 8 seems to make little difference. Should the span be syntactic (sentence or phrase rather than count of words)? Should the effect of a word on a decision be weighted inversely by its distance? I haven't coded such choices yet.

Another question is whether or not the results should "settle". At present, in all runs of the program, it has merely looked at each word once, comparing each sense of a word with all the definitions of every other word. But once the correct sense of one word is known, shouldn't only that sense's definition be used to evaluate the choice of other word senses? There are several ways in which such a settling might be done: these include left to right, or best guess first. It is possible that depending on the first choice made, other choices might be different; as suggested above, two separately consistent ways of assigning the sense definitions might be a metaphor. In the excerpt from *Moby Dick* above, when Melville wrote *your insular city of the Manhattoes* did he mean *insular* merely as "surrounded by water" or as "narrow-minded?" Of course he meant both. Melville, as a great author, used one word to convey two ideas, as opposed to the typical scientific paper which can go for pages without conveying any ideas at all. And this raises the possibility that one could automatically find metaphors by looking for texts that had two different semantic readings, each internally consistent. That is the most exciting future possibility of this work.

References

1. J. F. Sowa, *Conceptual Structures*, Addison-Wesley, 1984.
2. R. Granger, "Scruffy Text Understanding," *Proc. 20th ACL Meeting*, pp. 157-160, 1982.
3. R. Amsler and D. Walker, "The Use of Machine-Readable Dictionaries in Sublanguage Analysis," in *Sublanguage: Description and Processing*, ed. R. Grishman and R. Kittredge, Lawrence Erlbaum, 1985.

This basic scheme was outlined in the 1950s by Margaret Millar, and then again by Lawrence Urdang in the 1960s.

The cooperation of Oxford University Press, William Collins Sons & Co. Ltd, Merriam-Webster, the Oxford Text Archive, and the University of Waterloo Centre for the New OED is gratefully acknowledged.