# INFO-Y004: Natural Language Processing
# Assignment 3: Word Sense Disambiguation

Brian Delhaisse

June 6, 2015

## Contents



## 1 Introduction

The goal of this assignment is twofold: first, to simulate manually the original Lesk word overlap disambiguation algorithm, and second, to implement this algorithm using WordNet. The sentence to disambiguate is "*Time flies like an arrow*" in which each word should be disambiguated one at a time, from left to right, and in which earlier decisions are used later in the process.

For this assignment, I used Python (2.7.5) as the programming language, and the *Natural Language Toolkit* (*nltk*) library.

## 2 Manual Task

### 2.1 List of senses

The following definitions and examples comes from wiktionary [1]. The stop word 'an' has been removed from the sentence, and only the correct grammatical part (noun, verb,...) of the word definition is considered here for simplicity (I supposed that a (chart) parser has already analyzed the sentence):

[1] http://en.wiktionary.org/

- **time**:

  1. The inevitable progression into the future with the passing of present events into the past.
     *Time stops for nobody. the ebb and flow of time.*
  2. A duration of time.
     (a) A quantity of availability of duration.
     *More time is needed to complete the project. You had plenty of time, but you waited until the last minute. Are you finished yet? Time's up!*
     (b) A measurement of a quantity of time; a numerical or general indication of a length of progression.
     *a long time; Record the individual times for the processes in each batch. Only your best time is compared with the other competitors. The algorithm runs in O(n2) time.*
     (c) The serving of a prison sentence.
     *The judge leniently granted a sentence with no hard time. He is not living at home because he is doing time.*
     (d) An experience.
     *We had a wonderful time at the party.*
     (e) An era; the current era, the current state of affairs.
     *Roman times; the time of the dinosaurs*
     (f) A person's youth or young adulthood, as opposed to the present day.
     *In my time, we respected our elders.*
     (g) Time out; temporary, limited suspension of play.
  3. An instant of time.
     (a) How much of a day has passed; the moment, as indicated by a clock or similar device.
     *Excuse me, have you got the time? What time is it, do you guess? Ten o'clock? A computer keeps time using a clock battery.*
     (b) A particular moment or hour; the appropriate moment or hour for something.
     *it's time for bed; it's time to sleep; we must wait for the right time; it's time we were going.*
     (c) A numerical indication of a particular moment.
     *at what times do the trains arrive?; these times were erroneously converted between zones*
     (d) An instance or occurrence.
     *When was the last time we went out? I don't remember. see you another time; that's three times he's made the same mistake Okay, but this is the last time. No more after that!*
     (e) Closing time.
     *Last call: it's almost time.*
     (f) The hour of childbirth.
  4. The measurement under some system of region of day or moment.
     *Let's synchronize our watches so we're not on different time.*
  5. Ratio of comparison.
     *your car runs three times faster than mine; that is four times as heavy as this.*
  6. Tense.
     *the time of a verb*

7. The measured duration of sounds; measure; tempo; rate of movement; rhythmical division.
   *common or triple time; the musician keeps good time.*

- **flies (fly)**:

  1. To travel through the air, another gas or a vacuum, without being in contact with a grounded surface.
     *Birds of passage fly to warmer regions as it gets colder in winter. The Concorde flew from Paris to New York faster than any other passenger airplane. It takes about eleven hours to fly from Frankfurt to Hongkong. The little fairy flew home on the back of her friend, the giant eagle.*
  2. To flee, to escape (from).
     *Fly, my lord! The enemy are upon us!*
  3. To cause to fly (travel or float in the air): to transport via air or the like.
     *Charles Lindbergh flew his airplane The Spirit of St. Louis across the Atlantic ocean. Why don't you go outside and fly kites, kids? The wind is just perfect. Birds fly their prey to their nest to feed it to their young. Each day the post flies thousands of letters around the globe.*
  4. To be accepted, come about or work out.
     *Let's see if that idea flies. You know, I just don't think that's going to fly. Why don't you spend your time on something better?*
  5. To travel very fast.
  6. To move suddenly, or with violence; to do an act suddenly or swiftly.
     *a door flies open; a bomb flies apart*
  7. To hunt with a hawk.

- **like**:

  1. as, the way.
     *Winston tastes good like a cigarette should. Like you'd have them, like you'd have them, do unto you.*
  2. as if; as though.
     *It looks like you've finished the project. It seemed like you didn't care.*

- **arrow**:

  1. A projectile consisting of a shaft, a point and a tail with stabilizing fins that is shot from a bow.
  2. A sign or symbol used to indicate a direction.
  3. A directed edge.
  4. A dart.

## 2.2 Simulation of the original Lesk algorithm

Compared to the simplified Lesk algorithm, "the original Lesk algorithm is slightly more indirect. Instead of comparing a target word's signature with the context words, the target signature is compared with the signatures of each of the context words. In general, Simplified Lesk seems to work better than orginal Lesk". [1]

For the first word *"time"*, we compare the signature of each of its senses with the signature of each sense of each other word in the sentence. While comparing, we do not take into account the stop words such as 'the', 'an', 'a', 'I', 'which', 'that', etc. Once we

have found the sense that has the most common words with the other words signature, that is the sense that has the higher score, we considered it to be the sense of the word, and we go to the next word in the sentence (here, *"flies"*).

For the 2nd word *"flies (fly)"*, we do the same as the first word, but we only take into account the best sense that has been chosen for the previous word. We can see here that earlier decisions are important because they will affect our future choices. Once we have found the best sense for the 2nd word, we continue this process until the end of the sentence.

If there is a tie, the first sense is taken because they are normally ordered by sense that are the most common. The best sense for each word that has been found by simulating manually this original Lesk algorithm is:

- **time**: An instant of time (How much of a day has passed; the moment, as indicated by a clock or similar device).
- **fly**: To cause to fly (travel or float in the air): to transport via air or the like.
- **like**: As, the way.
- **arrow**: A sign or symbol used to indicate a direction.

For example, the best sense for the word "arrow" is the 2nd one because the verbs "use" and "indicate" are present in the signature of the best sense for the word "time", and no other words present in the other senses for the words "arrow" are overlapping with the rest.

# 3 Implementation

## 3.1 Code

To launch the code, just type the command *python assignment3.py* into the console.

In my code, I have 4 functions:

1. the *preprocess_sentence(sentence)* function which preprocesses the sentence by removing the punctuation, converting the uppercase characters into lowercase, and removing the stop words. The stop words were imported from the *nltk.corpus*, and contains words such as "*the*", "*an*", etc. This function returns a list of tokens.

2. the *simplified_lesk(word, sentence)* function which was implemented from the algorithm for the simplified Lesk algorithm described in the book [1] and in the course. It takes two arguments: *word* which is the word for which we need to find the best sense, and *sentence* (*-word*) which represents the context. It returns the best sense for the word passed in argument. This function has been implemented because according to the authors[1], "Simplified Lesk seems to work better than orginal Lesk".

3. the *original_lesk(sentence)* function which takes as argument the sentence. It performs the original Lesk algorithm on it, and returns a dictionary in which the keys are the words in the sentence, and the values are the best sense found for the corresponding word.

4. the *print_senses(word)* function which prints in the standard output the senses (definition and examples) of the word passed in argument.

The code can be found in the appendix.

### 3.2 List of senses

Here are the senses present in WordNet:

- **time**:
    1. an instance or single occasion for some event.
       *['this time he succeeded', 'he called four times', 'he could do ten at a clip']*
    2. a period of time considered as a resource under your control and sufficient to accomplish something.
       *['take time to smell the roses', "I didn't have time to finish", 'it took more than half my time']*
    3. an indefinite period (usually marked by specific attributes or activities).
       *['he waited a long time', 'the time of year for planting', 'he was a great actor in his time']*
    4. a suitable moment.
       *['it is time to go']*
    5. the continuum of experience in which events pass from the future through the present to the past.
    6. a person's experience on a particular occasion.
       *['he had a time holding back the tears', 'they had a good time together']*
    7. a reading of a point in time as given by a clock.
       *['do you know what time it is?', "the time is 10 o'clock"]*
    8. the fourth coordinate that is required (along with three spatial dimensions) to specify a physical event.
    9. rhythm as given by division into parts of equal duration.
    10. the period of time a prisoner is imprisoned.
        *['he served a prison term of 15 months', 'his sentence was 5 to 10 years', 'he is doing time in the county jail']*
    11. measure the time or duration of an event or action or the person who performs an action in a certain period of time.
        *['he clocked the runners']*
    12. assign a time for an activity or event.
        *['The candidate carefully timed his appearance at the disaster scene']*
    13. set the speed, duration, or execution of.
        *['we time the process to manufacture our cars very precisely']*
    14. regulate or set the time of.
        *['time the clock']*
    15. adjust so that a force is applied and an action occurs at the desired time.
        *['The good player times his swing so as to hit the ball squarely']*

- **flies**:
    1. (theater) the space over the stage (out of view of the audience) used to store scenery (drop curtains).
    2. two-winged insects characterized by active flight.
    3. flap consisting of a piece of canvas that can be drawn back to provide entrance to a tent.
    4. an opening in a garment that is closed by a zipper or by buttons concealed under a fold of cloth.

5. (baseball) a hit that flies up in the air.
6. fisherman's lure consisting of a fishhook decorated to look like an insect.
7. travel through the air; be airborne.
   *['Man cannot fly']*
8. move quickly or suddenly.
   *['He flew about the place']*
9. operate an airplane.
   *['The pilot flew to Cuba']*
10. transport by aeroplane.
    *['We fly flowers from the Caribbean to North America']*
11. cause to fly or float.
    *['fly a kite']*
12. be dispersed or disseminated.
    *['Rumors and accusations are flying']*
13. change quickly from one emotional state to another.
    *['fly into a rage']*
14. pass away rapidly.
    ***['Time flies like an arrow']**, 'Time fleeing beneath him']*
15. travel in an airplane.
    *[u'she is flying to Cincinnati tonight', 'Are we driving or flying?']*
16. display in the air or cause to float.
    *['fly a kite', 'All nations fly their flags in front of the U.N.']*
17. run away quickly.
    *['He threw down his gun and fled']*
18. travel over (an area of land or sea) in an aircraft.
    *['Lindbergh was the first to fly the Atlantic']*
19. hit a fly.
20. decrease rapidly and disappear.
    *['the money vanished in las Vegas', 'all my stock assets have vaporized']*

- **like**:

1. a similar kind.
   *['dogs, foxes, and the like", "we don't want the likes of you around here"]*
2. a kind of person.
   *["We'll not see his like again", "I can't tolerate people of his ilk"]*
3. prefer or wish to do something.
   *['Do you care to try this dish?', 'Would you like to come along to the movies?']*
4. find enjoyable or agreeable.
   *['I like jogging', 'She likes to read Russian novels']*
5. be fond of.
   *['I like my nephews']*
6. feel about or towards; consider, evaluate, or regard.
   *["How did you like the President's speech last night?"]*
7. want to have.
   *["I'd like a beer now!"]*

8. resembling or similar; having the same or some of the same characteristics; often used in combination.
   *['suits of like design', 'a limited circle of like minds', 'members of the cat family have like dispositions', 'as like as two peas in a pod', 'doglike devotion', 'a dreamlike quality']*
9. equal in amount or value.
   *['like amounts', 'equivalent amounts', 'the same amount', 'gave one six blows and the other a like number', 'the same number']*
10. having the same or similar characteristics.
    *['all politicians are alike', 'they looked utterly alike', 'friends are generally alike in background and taste']*
11. conforming in every respect.
    *['boxes with corresponding dimensions', 'the like period of the preceding year']*

- **arrow**:

1. a mark to indicate a direction or relation.
2. a projectile with a straight thin shaft and an arrowhead on one end and stabilizing vanes on the other; intended to be shot from a bow.

### 3.3 Results

For the *original* Lesk algorithm, the best sense found for each word is:

- **time**: a period of time considered as a resource under your control and sufficient to accomplish something.
- **flies**: pass away rapidly.
- **like**: prefer or wish to do something.
- **arrow**: a mark to indicate a direction or relation.

For the *simplified* Lesk algorithm, the best sense found for each word is:

- **time**: an instance or single occasion for some event.
- **flies**: pass away rapidly.
- **like**: a kind of person.
- **arrow**: a mark to indicate a direction or relation.

According to the documentation of the nltk library, the senses are ordered by frequency, that is, the first member of the list is the primary (most frequent) sense. Thus if there is a tie, the sense that is considered is the sense that is in front in the list.

## 4 Differences and overlaps between the manual and automated result

It's clear that the results depends on the dictionary we are using (see table 1). The senses found manually and automatically for the words "time" and "arrow" are overlapping (that is, they have some words in common, and are very similar). In the other hand, the senses found for the words "flies" and "like" are quite differents: The correct sense for the word "flies" would be the one found automatically, and the correct sense for the word "like" would be the one found manually. The difference for the word "flies", for example, can be

| Best Sense | Manually | Automatically |
|---|---|---|
| time | An instant of time (How much of a day has passed; the moment, as indicated by a clock or similar device) | a period of time considered as a resource under your control and sufficient to accomplish something. |
| flies | To cause to fly (travel or float in the air): to transport via air or the like. | pass away rapidly. |
| like | As, the way. | prefer or wish to do something. |
| arrow | A sign or symbol used to indicate a direction. | a mark to indicate a direction or relation. |

Table 1: Manual and automated results for the sentence "Time flies like an arrow".

explained by the fact that the corresponding definition in the wiktionary for "pass away rapidly" is "To travel very fast", but this last one contains no examples which reduces the size of its signature (that is, it contains less unique words to compare with). Concerning the word "like", knowing that it was a conjunction helped to find manually the correct definition. It shows that parsing the sentence before doing sense disambiguation can help and improve this last one.

## 5 Summary

This assignment allowed me to implement one of the algorithm which is used to disambiguate sentences. As written in the book [1], "The primary problem with either the original or simplified approaches, however, is that the dictionary entries for the target words are short, and may not provide enough chance of overlap with the context". Because of this, generally, the Corpus Lesk algorithm is used and is "the best-performing of all the Lesk variants" [1].

## References

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 20.4.1. Pearson Prentice Hall, 2008.

# Appendix - Code

```
# Assignment 3: Word Sense Disambiguation
# author: Brian Delhaisse
# Python 2.7.5
# references: course

from nltk.corpus import wordnet as wn
from nltk.corpus import stopwords

# Preprocessing: remove punctuation, put the sentence in lower case, and remove the stop words from the sentence.
# @return a list of words
def preprocess_sentence(sentence):
  import string
  sentence = ''.join(word for word in sentence if word not in tuple(string.punctuation))
  sentence = sentence.lower()
  stop = stopwords.words('english')
  return [i for i in sentence.split() if i not in stop]

# Compute the SIMPLIFIED Lesk algorithm (as described in the book "Speech and Language Processing" by Daniel Jurafsky & James H. Martin, 2nd edition, chap 20, section 4.1)
def simplified_lesk(word, sentence):
  senses = wn.synsets(word)
  best_sense = senses[0].definition() # The first member of the list is the primary (most frequent) sense. See documentation.
  max_overlap = 0
  context = set(sentence.split() if isinstance(sentence, str) else sentence) - set(word)

  # Compute the overlapping
  def compute_overlap(signature):
    val = 0
    for w in signature:
      if w in context: #O(1)
        val = val + 1
    return val

  for sense in senses:
    signature = sense.definition().split() # for definitions
    for example in sense.examples(): # for examples
      signature = signature + example.split()
    signature = set(signature)
    overlap = compute_overlap(signature)
    if overlap > max_overlap:
      max_overlap = overlap
      best_sense = sense.definition()

  return best_sense

# Compute the ORIGINAL Lesk algorithm
def original_lesk(sentence):
  sentence = preprocess_sentence(sentence)
  context = {w:{} for w in sentence}
  for w in sentence:
    for sense in wn.synsets(w):
```

```
51        context[w][sense] = preprocess_sentence(sense.definition()) # for
              definition
52        for example in sense.examples(): # for examples
53          context[w][sense] = context[w][sense] + preprocess_sentence(example)
54        context[w][sense] = set(context[w][sense])
55
56    for word in sentence:
57      senses = wn.synsets(word)
58      max_overlap = 0
59      best_sense = senses[0]
60      dico = {sense:0 for sense in senses}
61      for otherWord in sentence:
62        if word != otherWord:
63          for sense in senses:
64            for w1 in context[word][sense]:
65              for otherSense in context[otherWord].keys():
66                if w1 in context[otherWord][otherSense]: #0(1)
67                  dico[sense] = dico[sense] + 1
68
69      # finding the best sense
70      for sense in senses:
71        if dico[sense] > max_overlap:
72          max_overlap = dico[sense]
73          best_sense = sense
74
75      # removing other senses
76      for sense in senses:
77        if sense != best_sense:
78          context[word].pop(sense)
79
80    # computing a dictionary
81    dico = {word: context[word].keys()[0].definition() for word in sentence}
82    return dico
83
84  def print_senses(word):
85    senses = wn.synsets(word)
86    for sense in senses:
87      print sense.definition()
88      print sense.examples()
89
90  # Main code
91  sentence = "Time flies like an arrow."
92  sent = preprocess_sentence(sentence)
93  print 'SIMPLIFIED LESK: \n '
94  for word in sent:
95    print 'for the word \'%s\':' % word
96    #print print_senses(word)
97    print 'the best sense is: %s \n' % simplified_lesk(word, sent)
98
99  print '============================================================'
100 print 'ORIGINAL LESK: \n '
101 dico = original_lesk(sentence)
102 for w in sent:
103   print 'for the word \'%s\':' % w
104   print 'the best sense is: %s \n' % dico[w]
```