

LING 520: Computational Analysis of English

Semester: FALL '16

Instructor: Sowmya Vajjala

Iowa State University, USA

15 September 2016

Class outline

- ▶ Edit distance, Dynamic programming and spelling correction
- ▶ Morphology: overview
- ▶ Practice exercises

Edit Distance

- ▶ What is edit distance between words? what are edit operations?

Edit Distance

- ▶ What is edit distance between words? what are edit operations?
- ▶ What is minimum edit distance?

Edit Distance

- ▶ What is edit distance between words? what are edit operations?
- ▶ What is minimum edit distance?
- ▶ What is the minimum edit distance between google and goggle?

Edit Distance

- ▶ What is edit distance between words? what are edit operations?
- ▶ What is minimum edit distance?
- ▶ What is the minimum edit distance between google and goggle?
- ▶ What is the minimum edit distance between sleep and slept?

Edit Distance

- ▶ What is edit distance between words? what are edit operations?
- ▶ What is minimum edit distance?
- ▶ What is the minimum edit distance between google and goggle?
- ▶ What is the minimum edit distance between sleep and slept?
- ▶ How do we estimate the minimum edit distance between words?

How do we estimate edit distance?

- ▶ As it turns out, there are multiple ways of doing it. Let us take the same two words: sleep and slept.
- ▶ I can delete the second e, insert a t after p, to convert sleep to slept.

How do we estimate edit distance?

- ▶ As it turns out, there are multiple ways of doing it. Let us take the same two words: sleep and slept.
- ▶ I can delete the second e, insert a t after p, to convert sleep to slept.
- ▶ Or I can substitute the second e with p, and p with t.

How do we estimate edit distance?

- ▶ As it turns out, there are multiple ways of doing it. Let us take the same two words: sleep and slept.
- ▶ I can delete the second e, insert a t after p, to convert sleep to slept.
- ▶ Or I can substitute the second e with p, and p with t.
- ▶ There are only two real options here, but when you take longer words, and words of unequal length, possibilities become more and more.

How do we estimate edit distance?

- ▶ As it turns out, there are multiple ways of doing it. Let us take the same two words: sleep and slept.
- ▶ I can delete the second e, insert a t after p, to convert sleep to slept.
- ▶ Or I can substitute the second e with p, and p with t.
- ▶ There are only two real options here, but when you take longer words, and words of unequal length, possibilities become more and more.
- ▶ If we want, it is possible to create infinite ways of achieving the word to word conversions between any two words (even the above example).

How do we estimate edit distance?

- ▶ Humans can intelligently discard a few paths, and choose the best path of edits. Computers cannot.
- ▶ But computers can efficiently explore multiple paths simultaneously and reach a conclusion quickly.
- ▶ Dynamic Programming (DP) is one method that makes this process efficient.
- ▶ The idea of DP is to break a problem into a sequence of sub-problems, where solving each sub-problem will solve the next one.

- ▶ Let me take my sleep-slept example again. To transform sleep to slept, I go from left to right, looking at each character, and comparing with the target character.
- ▶ Although there can be several possibilities, some comparisons remain common between possibilities (e.g., s in source being s in target, l in source being l etc).
- ▶ If we store such common transitions, we don't have to calculate them again and again for each path, and we can use these numbers to get overall number of transitions for a given path.
- ▶ This can be visualized using a two-dimensional matrix with one word shown in rows, one word in columns.

		S	L	E	E	P
	0	1	2	3	4	5
S	1					
L	2					
E	3					
P	4					
T	5					

		S	L	E	E	P
	0	1	2	3	4	5
S	1	0	1	2	3	4
L	2	1	0	1	2	3
E	3	2	1	0	1	2
P	4	3	2	1	2	1
T	5	4	3	2	3	2

		S	L	E	E	P
	0	1	2	3	4	5
S	1	0	1	2	3	4
L	2	1	0	1	2	3
E	3	2	1	0	1	2
P	4	3	2	1	2	1
T	5	4	3	2	3	2

Pen and paper exercise

Following the approach described just now, find the edit distance between google and goggles.

Use for spelling correction/suggestions

What you saw just now is an edit distance known as Levenshtein distance, and is used to suggest spelling alternatives, by choosing the closest words to the mis-spelt word. How we detect misspellings is for another day.

Norvig's Spell Checker

- ▶ How many people went back and had a look at Norvig's spell checker article and code?

Norvig's Spell Checker

- ▶ How many people went back and had a look at Norvig's spell checker article and code?
- ▶ How many actually managed to run his code without getting errors (or how did you fix errors?)

Norvig's Spell Checker

- ▶ How many people went back and had a look at Norvig's spell checker article and code?
- ▶ How many actually managed to run his code without getting errors (or how did you fix errors?)
- ▶ Did you make any changes to make it run interactively?

Spelling suggestions in context

- ▶ What does context sensitive spell checking mean?

Spelling suggestions in context

- ▶ What does context sensitive spell checking mean?
- ▶ How do you think one can do context sensitive spell check?

Spelling suggestions in context

- ▶ What does context sensitive spell checking mean?
- ▶ How do you think one can do context sensitive spell check?
- ▶ methods: ngram approaches, grammar checking rules etc.
(More when we discuss NLP for CALL)

Additional readings/lectures on this topic

- ▶ Chapter on Writers Aids in Language and Computers by Dickinson et.al.
slides: <http://cl.indiana.edu/~md7/16/245/slides/02-writers-aids/slides.pdf>
- ▶ Lecture 2.5 on edit distance in Radev's coursera course, Chapter 3.10-3.11 in J&M.

Morphology for NLP: Quick Overview

- ▶ Morphological analysis is an important component in speech and language processing.
- ▶ Plays an important role for web search (capturing all morphological variants of a word usage, for example)
- ▶ Useful also in machine translation
- ▶ What is the big deal about morphological processing for NLP? If we have all word forms possible for all words, isn't it just a plain dictionary lookup?

Morphemes

- ▶ Morphemes: minimal, meaning-bearing units of language.
- ▶ Stem: main morpheme of the word
- ▶ Affixes: morphemes that add additional meanings or information to stems.
- ▶ cars is a word with two morphemes - car (stem) and -s (affix)
- ▶ Affixes: prefixes, suffixes, infixes (middle of the word), circumfixes (start and end of the word).
- ▶ clitic: a morpheme that is syntactically a word, but used in a reduced form with another word.

Combining Morphemes

4 ways: inflection, derivation, compounding, cliticization

- ▶ inflection: combining a stem with a grammatical morpheme, usually resulting in a word with same POS class (tag-tagged; car-cars)
- ▶ derivation: combining a stem with a grammatical morpheme, usually resulting in a word with different POS class (derive-derivation; computer-computerize)
- ▶ compounding: combining multiple word stems (greenhouse, redhead)
- ▶ cliticization: combining words with a clitic (we've, I'm)

Additional Readings/Lectures on this topic

- ▶ Survey of (mostly) English morphology (Chapter 3.1 in J&M)
- ▶ Lectures 2.01 to 2.04 in Radev's coursera course.
- ▶ I will continue with this topic on tuesday, and talk about Stemming and Lemmatization

Next Week

- ▶ Topics: Morphological analysis - stemming and lemmatization, Introduction to n-gram approaches.
- ▶ Readings: Chapter 3–4 in J &M, Chapter 5 in NLTK Book

Practice exercises

1. Figure out whether NLTK has a distance metric such as Levenshtein or other such orthographic distances, and learn how to use one such measure to get distance between words.
2. Check for any python based spell checking libraries. If you do not find any, learn to use PyEnchant library for spell checking.

Practice exercises

1. Figure out whether NLTK has a distance metric such as Levenshtein or other such orthographic distances, and learn how to use one such measure to get distance between words.
2. Check for any python based spell checking libraries. If you do not find any, learn to use PyEnchant library for spell checking.
3. Start doing problems in Problem Set 2 (see Blackboard)