Syntax & Dep_Search Lecture

...

0. Dependency Parsing and This lecture

- Two dependency Parsers used in this course
  - Turku Pipeline for Finnish based on Mate-tools
  - Stanford Parser for English
- Both produce parses in the same dependency scheme and format
- As probably already mentioned the topic of this lecture is dependency trees
  - More specifically what purpose do they serve for this pursuit
- By now practicalities of dependency parsing have been covered during the demo sessions
  - And you probably already have some dependency parsed text
- Since the point of this lecture is you to learn, please ask whenever you feel like it!

1. Motivation for this talk

- So, what could you or the powers you represent benefit from these dependency parses?
- And how would you go about using these for text mining?

2. Dep_Search / SETS query tool

- Before we seek to answer the earlier questions, let's briefly introduce our tool for querying treebanks
- Treebank here means simply a collection of sentences with their dependency graphs
- Querying here means that we should be able to find the based on their syntactic features
- Since you might be using the software in the demo sessions, we will briefly demonstrate how it works

2. Dep_Search / SETS query tool

- Can be run on either command line or through a webUI
- Python / Cython / C++
- Scales very well up to corpora of billions of tokens
- Easy to be embedded into other software
- Tested on Linux and OSX
- Rich query language (meaning it's both simple and powerful)
  - Best shown through examples

http://bionlp-www.utu.fi/dep_search/

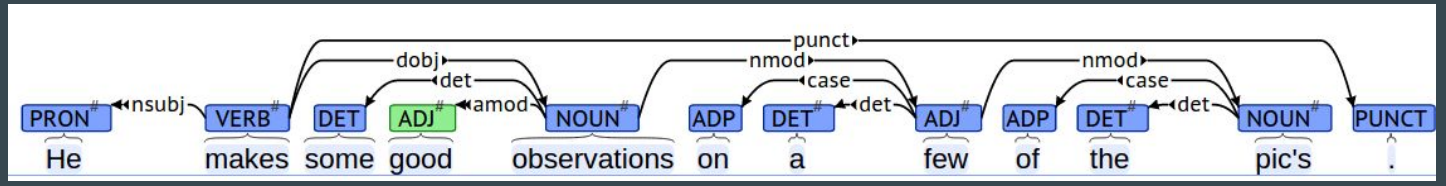
https://github.com/fginter/dep_search

Text Mining and Syntactic Analysis #1

- The main, unifying purpose of all text mining tasks is to from raw text produce information
  - This information can be for example:
    - Some kind of Sentiments
    - Entities or concepts
    - Relations between said entities
    - Etc
  - Anyway, the point could be said to be to automatically read and produce results of some kind from the text
  - In this pursuit knowledge is power, more the knowledge more the power
    - I guess all possible power and all possible knowledge would render this task redundant
  - Anyway, syntactic analysis gives us more knowledge of the text
    - Namely, its syntactic structure and information of the roles of the tokens in the sentence

1. Text Mining and Syntactic Analysis #2

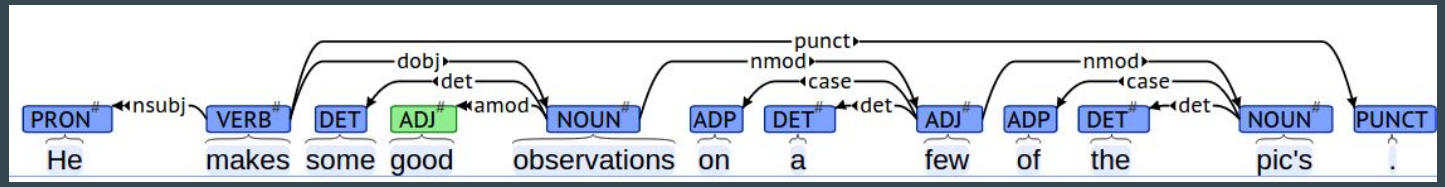
- Let's examine the knowledge we have of the sentence given by the dependency graph



- We see named relations between tokens.
- Which of them could be used to extract information and how
  - Aren't they in a way already information?

1. Text Mining and Syntactic Analysis #3

- Yes, we were examining this graph:



- We can at least see an adjective depending as an adjective modifier from a noun
- That's an easy piece of information to start off our information mining business!
- Let's see what kinds of stuff can we extract from our treebank with this very very simple idea

1. Text Mining and Syntactic Analysis #4

- Let's start our exploration from querying for adjectives from Finnish web-data using our query system
  - The search term here is simply "ADJ", the POS-tag name for adjective
  - http://bionlp-www.utu.fi/dep_search/?db=English&search=ADJ
  - http://bionlp-www.utu.fi/dep_search/?db=Finnish&search=ADJ
- Yep, our simple search really did find us some adjectives
  - But such query would be doable simply with grep tool and a POS-tagger, no big deal
- Let's include syntactic features into our query
  - Let's search for a subject with an adjective
  - http://bionlp-www.utu.fi/dep_search/?db=English&search=_+%3Camod+%28_+%3Cnsubj+_%29
- Now, this is not information, not yet. This is only of interest to linguists if even them!

1. Text Mining and Syntactic Analysis #5

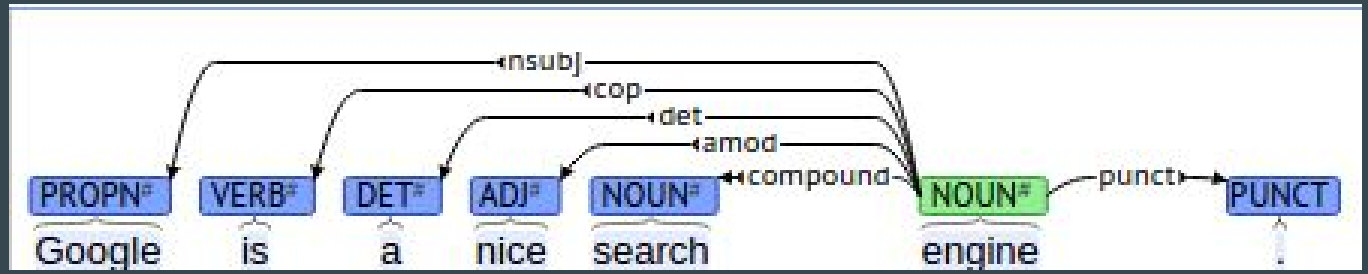
- With this simple understanding, can we gather any information?
  - Let us assume we represent cats, we are hired by cats to study with what adjectives they are described
  - And with which adjectives are their competitors; dogs described
  - We can with our current knowledge do a search
  - http://bionlp-www.utu.fi/dep_search/?db=Fi-Parsebank-1M&search=_+%3Camod+kissa
  - http://bionlp-www.utu.fi/dep_search/?db=Fi-Parsebank-1M&search=_+%3Camod+koira
  - (Finnish only, sorry couldn't find proper english example)

1. Text Mining and Syntactic Analysis #6

- On the other hand we could, with our current knowledge, search for things which are described by certain adjective
    - Let's say ugly and beautiful
    - http://bionlp-www.utu.fi/dep_search/?db=ukw&search=_+%3ENMOD+beautiful
    - http://bionlp-www.utu.fi/dep_search/?db=ukw&search=_+%3ENMOD+ugly
    - And in Finnish:
    - http://bionlp-www.utu.fi/dep_search/?db=Fi-Parsebank-1M&search=_+%3Eamod+kaunis
    - http://bionlp-www.utu.fi/dep_search/?db=Fi-Parsebank-1M&search=_+%3Eamod+ruma
- That's all very cute and fulfills the criteria of mining text
    - But one could still go about doing all that with regular expressions just fine
    - Why the trees?

1. Text Mining and Syntactic Analysis #7

- Let's move onto a little more complicated syntactic structures a miner of text might be interested in



- *Something* is *Something*
- "What are *they* thinking about us"
  - Much less ambiguous than the previous example
- We can see the subject and copula depending on a single node

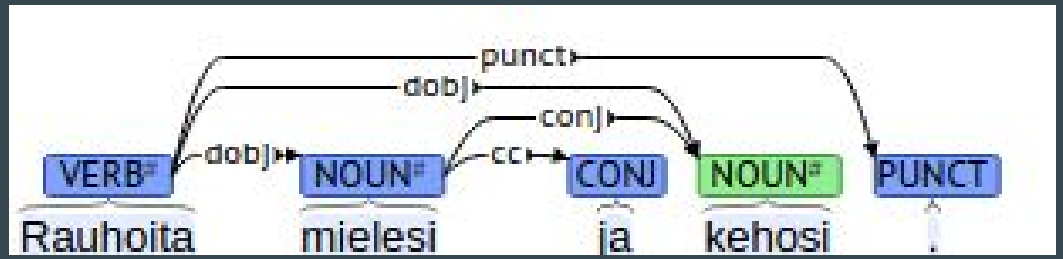
1. Text Mining and Syntactic Analysis #8

- Once again we formulate a query
  - http://bionlp-www.utu.fi/dep_search/?db=English&search=_+%3Ecop+_+%3Ensubj+_
  - Hmmm... Not quite as clean as we'd like. Let's add some restrictions a la google example

  - http://bionlp-www.utu.fi/dep_search/?db=English&search=NOUN+%3Ecop+_+%3Ensubj+PROPN
  - Much Better!
  - First two hits describe in a concrete way the actions of a large organization
  - These are the kind of examples somebody would be hired to harvest

1. Text Mining and Syntactic Analysis #9

- Somebody could also be hired to find stuff which is mentioned with our target
- Let's say we are tasked by just somebody to find out what entities are mentioned with the city of Moscow in Finnish internet data
- To do this we could exploit conjoining dependencies in the dependency trees
- They look like this:

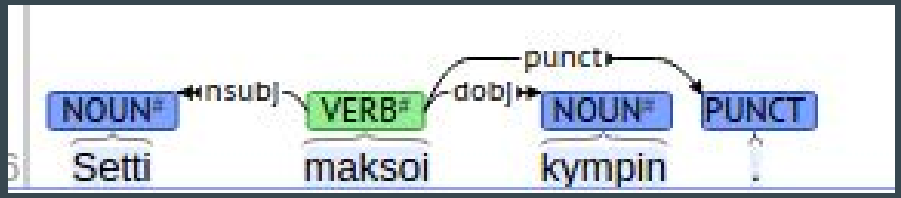


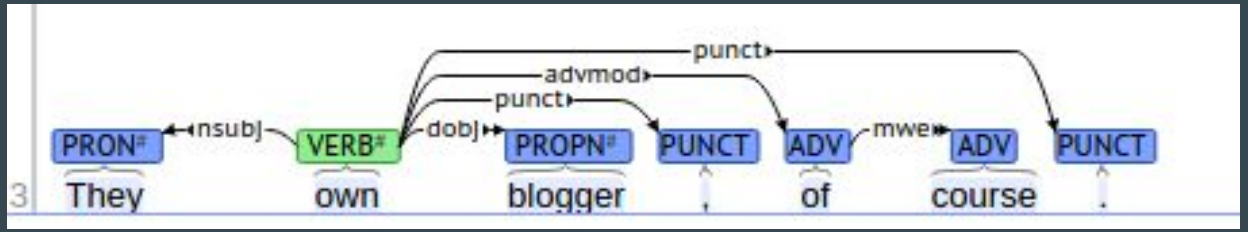
1. Text Mining and Syntactic Analysis #10

- So let us formulate a query in which the city of Moscow is asked to depend as a conjunction
    - Like this http://bionlp-www.utu.fi/dep_search/?db=Fi-Parsebank-1M&search=_+%3Cconj+L%3DMoskova
    - The results seem to make sense and are not too alarming
    - Once again a little nasty to do with regex, but works well with dependency trees

1. Text Mining and Syntactic Analysis #11

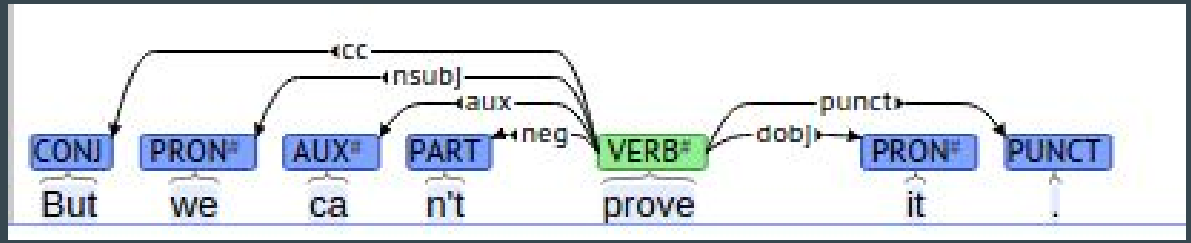
- Another simple and useful tangible thing we could extract from a collection of these dependency graphs is Subject - Verb - Object -- triplets
    - http://bionlp-www.utu.fi/dep_search/?db=English&search=VERB+%3Ensubj+_+%3Edobj_





1. Text Mining and Syntactic Analysis #12

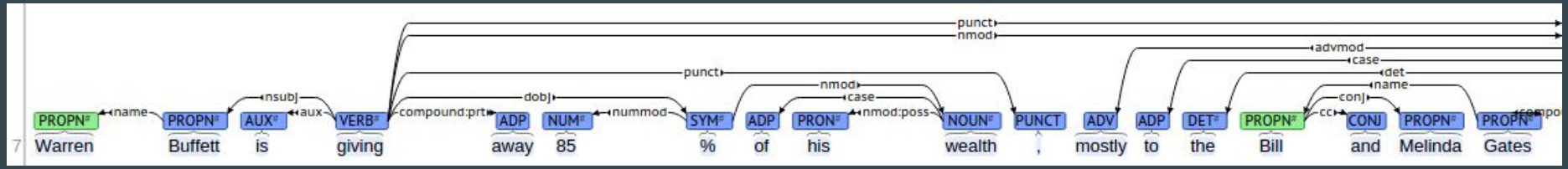
- But please be careful and check your data!



- Oh yeah, what can these be used for?
  - For example finding out a typical object for a given verb or subject-verb pair
  - What do people eat? Or something like that.

1. Text Mining and Syntactic Analysis #13

- Can we still come up with some simple examples?
- How about names? They have an explicit dependency relation, like this:
  - http://bionlp-www.utu.fi/dep_search/?db=English&search=_+%3Cname+_



- One could extract names of people generally or maybe something like names of people which are subjects to playing golf

2. More Complicated Uses

- These examples were all quite simple and used explicit information given by the dependency graphs
- And is only meant to be an introduction to the capabilities of the trees and also some features of the query system
- That doesn't mean that's all they are good for
    - Well, what else are they then good for?
- For example dependency trees can be used to generate features for machine learning, here mainly classification and clustering of text and tokens
- Paths in the dependency graphs can be used to find relations between tokens

3. Real Uses

- To showcase that I'm not just hyping these dependency graphs for the fun of it, let us have look at google scholar with the keywords text mining dependency trees
    - https://scholar.google.fi/scholar?start=40&q=text+mining+dependency+trees&hl=fi&as_sdt=0,5