How to write a master's thesis: a computational linguist's view

With an excursus on structural disambiguation with distributional methods

Simon Šuster
24/4/2014

Introduction

- My master's thesis is in natural language processing
- The problem of structural ambiguity and how I address it
- On the practical side, I'd go about writing my thesis differently now

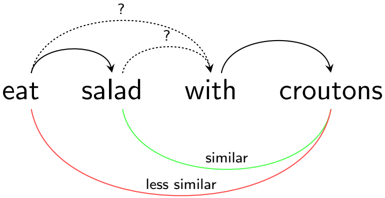
This talk, more emphasis on:

- advice based on my own mistakes and suggestions of others
- most points should be valid for any kind of research involving experimental/computational work
  - but adapt to your situation
  - and respect the formal thesis requirements

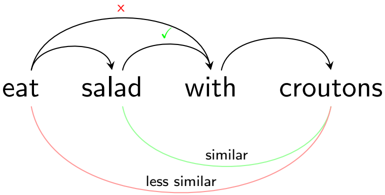
Shortly on my thesis

- Statistical syntactic parsers are frequently used to obtain dependency representation of sentences.
- Parsers perform differently well on different structural ambiguities.
- One of the most difficult cases is prepositional phrase attachment (PPA).
- Error rate on this type is above average for a data-driven parser for French.
- Idea: introduce an (external) component specialized in PPA for better decisions
  - investigated before, but we address the problem from a distributional-semantics perspective
  - semantic similarity between elements in an ambiguous PPA case indicates correct attachment

Shortly on my thesis



Shortly on my thesis



- Semantic similarity ⇒ sharing distributional properties ⇒ closeness between two word vectors
- Build vector-space models, also treat PP as a single unit/vector (by composition)

Shortly on my thesis

What we show:

- on a dataset of ambiguous cases, #correct > #incorrect attachments introduced by our method
- by pre-attaching PPs in a raw text, then running the parser, *improved* parsing accuracy (compared to no pre-annotation)

Downside:

- improvement is tiny (restricted space for improvement)
- semantic similarity only reliable for strongly related words

How to write a master’s thesis?

6 suggestions

Suggestion #1: Write the paper first

Paper is paramount, thesis is secondary

- a narrower, less intimidating object than thesis
- also, chances of ever publishing your work are much higher
    - and paper submission deadlines are a great motivation

The paper is thesis' skeleton around which you build details such as:

- further experimental information
- related work
- general introduction to the field
- description of less successful experimental attempts

Separate core from details

Throughout the research, keep at least two documents:

- one with all technical details of all experiments you run
  - together with ideas you'd like to put into practice
- the paper (∼8–10 pages), allows you to stay focused on main ideas

Thesis is then a carefully thought-out compilation of both documents.

Suggestion #2: Writing before experimenting

Start writing *soon*, before doing experiments (assuming a good research problem)

- better planning
- more focus
- crystallization of the problem
- easier to weed out unnecessary experiments

Start writing on:

- what is the problem and how you address it
- how you intend to run experiments
- why you run them, and what they will show
- (some) related work

Note: spending much time early on on writing literature review can block new views on the topic

Suggestion #3: Read a lot

Go to great lengths to find papers close to your research problem

- to prevent duplication of effort
- and because replication is only a very small contribution

Related work sometimes hard to find due to varying terminology

- Help from your supervisor, conference, mailing lists (e.g. corpora-list)

Suggestion #3: Read a lot

Go to great lengths to find papers close to your research problem

- to prevent duplication of effort
- and because replication is only a very small contribution

Related work sometimes hard to find due to varying terminology

- Help from your supervisor, conference, mailing lists (e.g. corpora-list)

Remember the idea of *compounding* (adapted from R. Hamming):

> The more you read, the more you know
> The more you know, the easier to read
> The more you know, the more you do
> The more you do, the more the productivity

“The reading is necessary to know what is going on and what is possible. But reading to get the solutions does not seem to be the way to do great research.” (R. Hamming)

Suggestion #4: Choose a topic with care

Aim for a small contribution to a big, unanswered problem

- *small* contribution, concrete plan

It's OK to work on a topic predetermined by your supervisor

- if you're attracted by the topic
- if he/she knows the topic background very well (normally so)

It's great to come up with a topic yourself, but:

- talk to people knowing the sub-field better than you do (includes your supervisor)

Suggestion #5: Tell a story

(Cf. talk by Simon Peyton Jones `http://youtu.be/g3dkRsTqdDA`)

Thesis/paper is a narrative:

- here is a problem
- it's an interesting problem
- it's an unsolved problem
- here is my idea 🔔
- my idea works: data, experiments
- relation to other work

Suggestion #6: Painstaking documentation

You want to be able to re-run the experiments several years on from now

- everything needs to be in a single place, with a single entry point such as a readme with instructions to proceed
- you don't want to be figuring out which data set produced a specific plot, or what parameters were used

Back-up frequently or use a versioning system for all your documents and code (e.g. free private Bitbucket accounts)

Further reading

The bulk of these guidelines pertain to master's students as well:

- advice on writing clear and concise sentences:

  http://homepages.inf.ed.ac.uk/sgwater/writing_advice.html

- finding research problems, how to read papers, and much more:

  http://www.cs.jhu.edu/~jason/advice/

- "How to be a successful PhD student":

  http://people.cs.umass.edu/~wallach/how_to_be_a_successful_phd_student.pdf

- "You and your research", by R. Hamming