UNIVERSITY OF CAMBRIDGE

Learning a Theory of Marriage (and other relations) from a Web Corpus

Sandro Bauer
Computer Laboratory, University of Cambridge
sandro.bauer@cl.cam.ac.uk

Motivation

- Long-term goal of NLP: Natural language understanding (just like humans do!)

- NLP has been very successful at interpreting individual words, sentences and documents:
  - POS tagging
  - Syntactic parsing
  - Semantic parsing
  - Word sense disambiguation
  - Co-reference resolution
  - ...

But texts contain much more information than is stored on the surface level

Can we infer some of this knowledge using NLP methods and make it available to the computer?

Idea

How do humans learn what a concept such as „marriage“ typically involves?

Various possibilities:
- Look it up in a dictionary
- Ask other people
- Wait for the right person and give it a try yourself
- Look at other married couples and observe what they're doing...

Computers can't yet fall in love,
but they can crawl texts quite well!

Let's have a look at the whole web and what it tells us about married couples!

Resources used

- **Web corpus** ClueWeb09: 2 TB compressed – a good portion of the WWW on disk
- Freebase (a state-of-the-art **knowledge base** with structured information about 2 million people)
- Lots of processors

Extraction of relations from the corpus

- Compile a list of all married couples in the knowledge base
- Crawl the web corpus and look for sentences which contain the names of two married people
- Extract the **relation** the two people are in
- Requires a lot of NLP machinery
  - Boilerplate removal
  - Splitting documents into sentences
  - Tokenisation
  - POS tagging
  - Dependency parsing
  - …

Output

`say sth.`
`tell sth.`
`see`
`get engaged to`
`file for divorce from`
`join`
`travel with`
`visit`
`marry`

Applying a weighting scheme

That looks great, but is this what we typically have in mind when we think about marriage?

Not quite. Not only married people meet and visit each other!

Aim: Work out which relations are specific to married couples
Idea: Use a parsed background corpus and look how many instances there are

Results:

Typical of marriage	Less typical
be someone's wife	hit someone
be engaged to	defend
date	remember
meet	turn to
divorce	rejoin
file for divorce from	promise sth.
tie knot with	bury
love	remind
expect child with	look at