(three) Ideas for master projects

Malvina Nissim
m.nissim@rug.nl
room 1311.421

IK meeting, 11 December 2015

Project 1: Authors and Genres

Authorship verification and genre detection

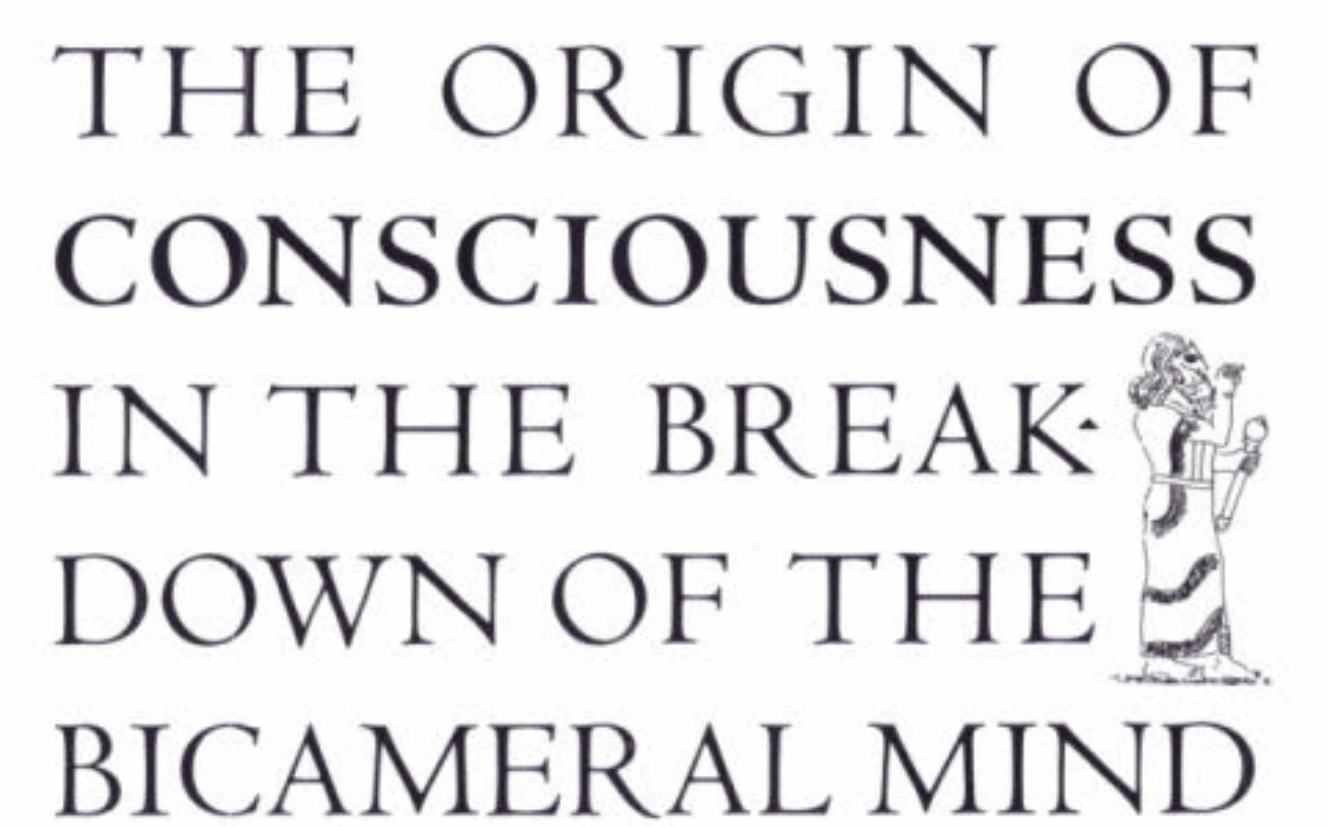
with Martijn

- authorship verification: given a document, tell whether it's been written by the same person who wrote a set of other known, given documents.
  - binary problem: not picking one of several authors, rather deciding whether same author or not
- genre detection: given a document, assign it to a given genre.
  - can be seen as multiclass classification problem, but also as binary, per genre
- **research**: mutual benefit? features?

Project 2: the Alternative project

Words of Consciousness

Julian Jaynes' controversial theory on the origin of consciousness



Words of Consciousness

What does it have to do with language?
[from Wikipedia]

Jaynes asserted that, until roughly the times written about in Homer's Iliad, humans did not generally have the self-awareness characteristic of consciousness as most people experience it today.

For example, in the Iliad and sections of the Old Testament no mention is made of any kind of cognitive processes such as introspection, and there is no apparent indication that the writers were self-aware. According to Jaynes, the older portions of the Old Testament (such as the Book of Amos) have few or none of the features of some later books of the Old Testament (such as Ecclesiastes) as well as later works such as Homer's Odyssey.

Words of Consciousness

What does it have to do with language processing?

- recent studies using LSA on documents from different ages show some evidence (Diuk et al 2012)
- recent studies to explore word contexts across time (Zhang et al 2015)

see document for more info.

Project 3: Winograd Schema Challenge

Winograd Schema Challenge

Although they ran at about the same speed, Sue beat Sally because she had such a [good/bad] start.
Who had a [good/bad] start?
Answers: Sue/Sally.

I spread the cloth on the table in order to [protect/display] it.
To [protect/display] what?
Answers: the table/the cloth.

Why it is difficult, and interesting

examples are constructed such as they are:

- easily disambiguated by the human reader (ideally, so easily that the reader does not even notice that there is an ambiguity)
- not solvable by simple techniques such as selectional restrictions
- Google-proof: there is no obvious statistical test over text corpora that will reliably disambiguate these correctly

Key points

- anaphora resolution
- learning approach, but deep processing
- can we exploit paraphrases? (available)
- can we exploit scores of association norms (available)?

Data and References

- Main attempt: overall accuracy of 73% using a composite set of features in a learning setting, on this dataset: `www.hlt.utdallas.edu/~vince/data/emnlp12/train-emnlp12.txt`.
  Altaf Rahman and Vincent Ng (2012), "Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge". In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea.

- Overview of datasets and links: `https://www.cs.nyu.edu/davise/papers/WS.html`
- Data on association norms: `https://sites.google.com/site/kenmcraelab/norms-data`