UNIVERSIDADE
BEIRA INTERIOR

SNLP: Statistical Natural Language Processing

Special Session along with ICIMP 2020
The Fifteenth International Conference on
Internet Monitoring and Protection
September 27, 2020 to October 01, 2020

Lisbon, Portugal

https://www.iaria.org/conferences2020/ICIMP20.html

Sebastião Pais - sebastiao@di.ubi.pt
&
Irfan Tanoli - irfan.khan.tanoli@ubi.pt

Chairs

Sebastião Pais

Sebastião Pais received the PhD degree from MINES ParisTech - PSL, Paris. He is currently a Professor at the Computer Science Department, the University of Beira Interior, and Researcher at NOVA-LINCS and GREYC Laboratory. His research and teaching interests are in the areas on artificial intelligence, statistical natural language processing, lexical semantics, machine learning and unsupervised and language-independent methodologies.

Irfan Khan Tanoli

IRFAN KHAN TANOL received the B.S degree in computer science from Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Karachi, Pakistan. Later he received the M.S degree in software engineering from Technical University of Madrid, Madrid, Spain. From 2015 to 2019, he was a PhD research student at Gran Sasso Science Institute, L'Aquila, Italy where he obtained his PhD degree in computer science in 2019.

Why study (statistical) NLP

- (Most of) you are studying in a 'computational linguistics' program
- Many practical applications
- Investigating basic questions in linguistics and cognitive science (and more)

Application examples ...

For profit (engineering)

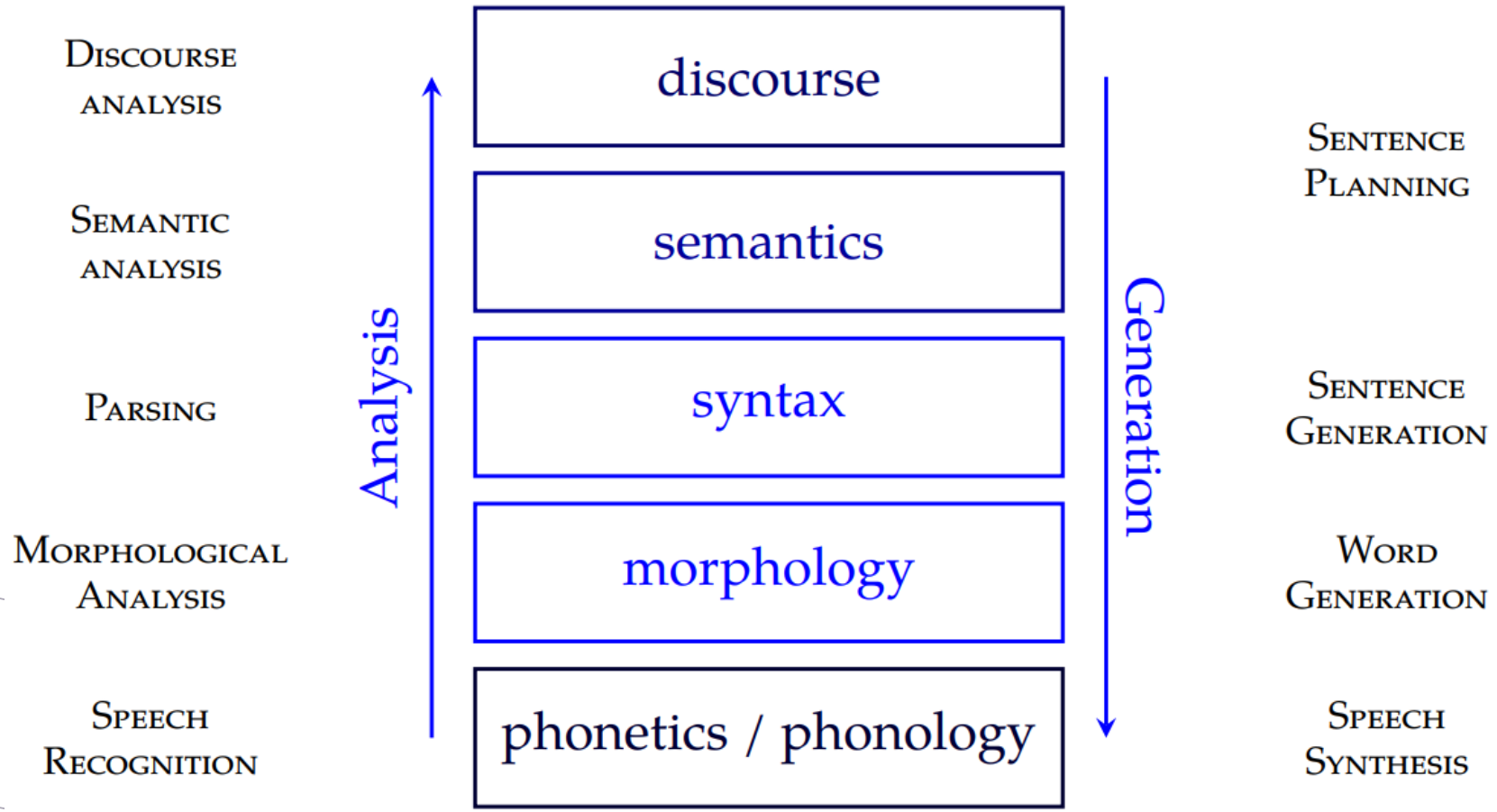
- Machine translation
- Question answering
- Information retrieval
- Dialog systems
- Summarization
- Text classification
- Text mining/analytics
- Sentiment analysis
- Speech recognition and synthesis
- Automatic grading
- Forensic linguistics
- ...

... Application examples

For fun (research)

- Modeling language processing learning
- Investigating language change through time and space
- (Aiding) language documentation through text processing
- (Automatic) corpus annotation for linguistic research
- Stylometry, author identification
- ...

Layers of linguistic analysis



On the word 'statistical'

But it must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term. — Chomsky (1968)

- Some linguistic traditions emphasize(d) use of 'symbolic', rule-based methods
- Some NLP systems are based on rule-based systems (esp. from 80's 90's)
- Virtually, all modern NLP systems include some sort of statistical component
- Language-Independent Approaches

What is statistical NLP?

Statistical NLP methods

- Involve deriving numerical data from text
- Are usually but not always probabilistic (broad church – we include e.g., vector spaces)

What is statistical NLP?

Relation to wider context

- Matches move from logic-based AI to probabilistic AI
  - Knowledge - probability distributions
  - Inference - conditional distributions
- Probabilities give opportunity to unify reasoning, planning, and learning, with communication
- There is now widespread use of machine learning (ML) methods in NLP
- Use of approximation for hard problems

Future Challenges

- Breaking sentences into tokens
- Tagging parts of speech (POS)
- Building an appropriate vocabulary
- Linking the components of a created vocabulary
- Understanding the context
- Extracting semantic meaning
- Named Entity Recognition (NER)
- Transforming unstructured data into structured data
- Ambiguity in speech
- Using Unlabeled Data for Translation
- Unsupervised Word Translation
- ...

ICIMP x/SNLP ...

A Lexicon Based Approach to Detect Extreme Sentiments

- Unsupervised lexicon-based approach that detects the extreme sentiments from social media
- Unsupervised approach for automatic detection of people's extreme sentiments on social networks
- Automatically to build a standard lexicon consisting of extreme sentiments terms having high extreme positive and extreme negative polarity
- Evaluated the performance on five different social networks and media datasets

... ICIMP x/SNLP ...

Language-Independent Approaches to Detect Extremism and Collective Radicalization Online

- Study about language-independent approaches to detect extremism and collective radicalization online
- Understanding and detecting extremism and collective radicalism on the online world has a connection to sentiment analysis and opinion mining
- The main focus of this work is to find the best ways to identify extremism and collective radicalisation on the internet using sentiment analysis focusing on statistical and probabilistic methods, in order to create an unsupervised and language-independent approach

... ICIMP x/SNLP

Modeling Natural Language Policies into Controlled Natural Language: A Twitter Case Study

- Social network providers usually describe the terms of data storage, usage, and sharing, by adopting natural languages
- Controlled Natural Languages are subsets of natural languages that are obtained by restricting the grammar and vocabulary, to minimize - or even eliminate - ambiguity and complexity of NL
- In this paper is study some policy-oriented controlled natural languages

Thanks for your presence.