Natural Languages and Semantics

Adam Przepiórkowski

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warsaw

CHIST-ERA 2014, Istanbul

18 June 2014

Human Language Understanding (HLU)

Human Language Understanding:

- natural languages (language-specific resources and methods),
- semantics and pragmatics (lexical and compositional).

SMT

Hürriyet (`http://www.hurriyet.com.tr/`), 16 June 2014:

Geçtiğimiz ay yoğun bunalımda olduğu, hatta sosyal medyada "intihar ettiği" konuşulan ünlü aktris Nicole Kidman, yepyeni bir imajla kamuoyunun karşısına çıktı. Kidman'ın artık yaşlandığı için yapımcı ve yönetmenlerin eskisi kadar ilgisini çekmediği bu yüzden de önemli filmler için teklif almadığı ve depresyonda olduğu yazılıyordu.

Via **Google Translate**:

In the past month crisis is intense, even in social media, "suicide" spoken famous actress Nicole Kidman, appeared before the public in a brand new imaging. Because the producer and director of Kidman's old now so old so important as attracting and getting offers for films that were written in depression.

Pendulum swung too far...

Last 20 years:

- the rule of **statistical paradigm in NLP**,
- emphasis on **language-independent approaches** (in research and, even more so, in funding).

Success story – **part-of-speech tagging**:

- language-independent **machine learning** methods,
- trained on relatively **small manually-annotated corpora**,
- over **97% accuracy** for English.

Other tasks also beneficial:

- word sense disambiguation,
- sentiment analysis,
- shallow and less shallow (e.g., dependency) parsing,
- to some extent: machine translation, etc.

Pendulum swung too far... (contd.)

Acknowledged **limits of this dominant paradigm**:

Chris Manning (2011; CICLing proceedings):

- subtitle: "**Is It Time for Some Linguistics?**",
- 97% word-level accuracy of best taggers for English translates into 56% of sentence-level accuracy,
- "I suggest... that the largest opportunity for further progress comes from improving the taxonomic basis of the linguistic resources from which taggers are trained. That is, from improved descriptive linguistics."

Ken Church (2011; LiLT):

- title: "**A Pendulum Swung Too Far**",
- "...we may have been too successful. Not only have we succeeded in making room for what we were interested in, but now there is no longer much room for anything else."

Thesis 1

The **pendulum swings back**...

Thesis 1: in order to advance human language understanding, **we need** to:

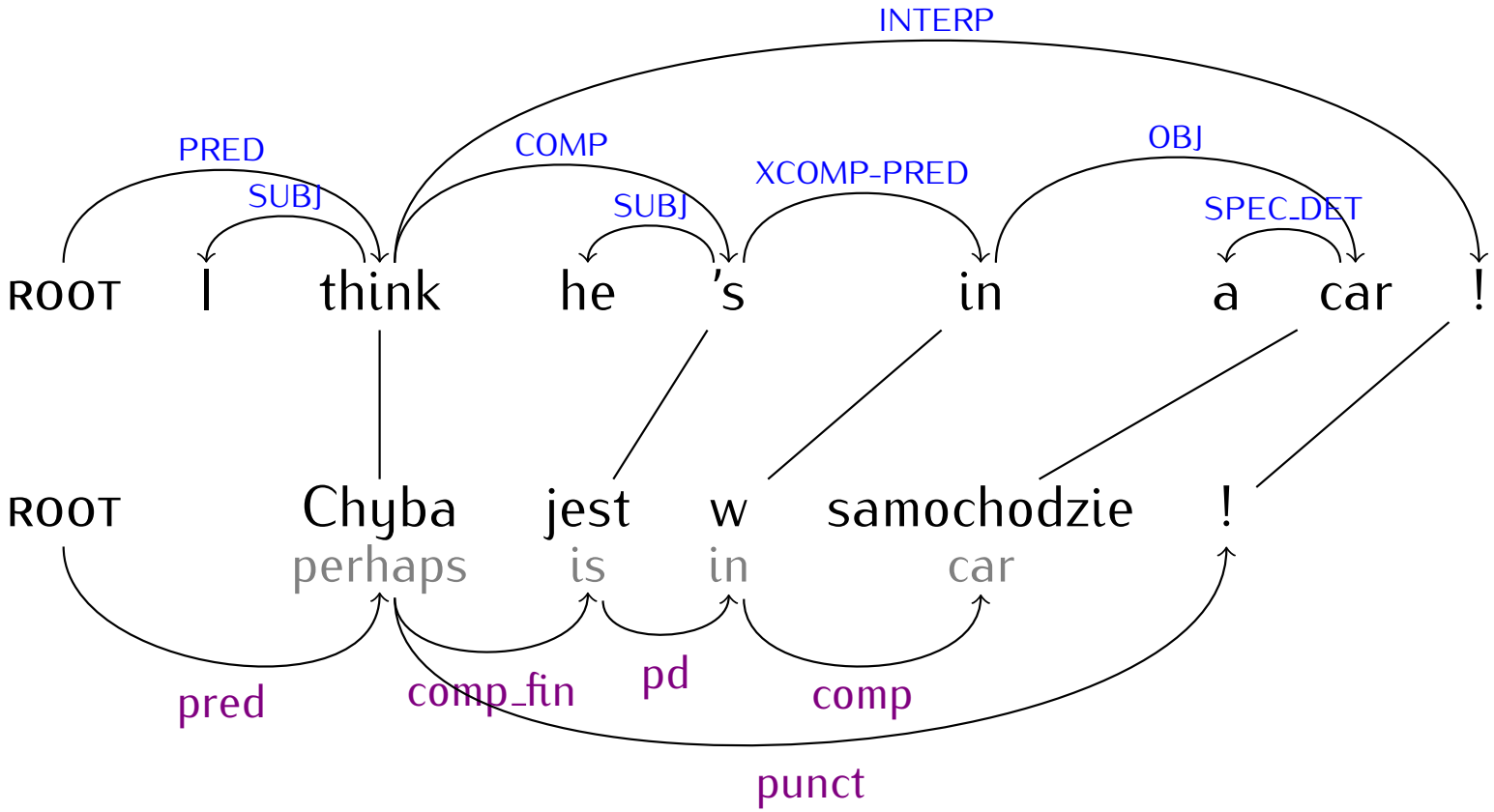
- put more emphasis on **language-specific resources and methods**,
- put more effort into **combining statistical** (largely "language-independent") **and symbolic** (largely "language-specific") **approaches**.

Example 1

Projection of grammatical information (Wróblewska and Przepiórkowski @ LREC 2014 and PolTAL 2014):

- input: a **large parallel English–Polish corpus**,
- parse the English part with the LFG grammar for English:
  - the **grammar manually developed by linguists**,
  - uses various **shallow techniques**,
  - probabilities of parses **trained on a treebank**,
- select **dependency-like information** (subject, object, etc.),
- **transfer it to Polish** sentences:
  - via **word alignment links**,
  - using **graph algorithms** for better transfer and for constructing Polish dependency trees,
  - with some **linguistic rules** rearranging resulting structures and translating labels.

Example 1 (contd.)



Results:

- a large dependency treebank for Polish,
- a wide-use dependency parser for Polish.

Understanding

Last 20 years – great progress in **lexical semantics**:

- **word sense disambugation**,
- semantic lexica, esp., **wordnets** (for many languages),
- **distributional semantics**:
  - meanings as co-occurrence vectors,
  - finding similar words, hyperonyms, etc.

Much less progress in **compositional semantics**:

- relatively large lexica or corpora of **semantic roles** mainly for English (FrameNet, VerbNet, PropBank),
- some **niche work** (mainly for English):
  - within categorial grammar (e.g., Boxer – DRT structures),
  - on underspecified approaches (e.g., within HPSG),
  - on linear logic (Glue Semantics in LFG).

Thesis 2

Thesis 2: in order to advance human language understanding, **we need** to:

- put more **emphasis on compositional semantics**,
- find ways to **combine** (largely statistical approaches to) **lexical semantics and** (largely – but not only – symbolic approaches to) **compositional semantics**.

One (not unreasonable) way to go: construct **compositional semantic resources** such as VerbNets and semantic grammars for various languages.

More exciting (and cutting-edge) research: **combine statistical and compositional approaches**, minimising human input.

Example 2

Lewis and Steedman 2013 (Transactions of ACL 1):

- title: "**Combined Distributional and Logical Semantics**",
- **manually constructed** lexical entries for a **small class** of words expressing **semantic operators**, e.g., *not*, *every*, *and*, etc.,
- lexical entries of **all content words automatically learned** – from a large corpus parsed with a parser trained on a treebank,
- **lexical entry**:
  - combinatory syntax (categorial grammar class),
  - semantics (corresponding first order logic lambda expression),
- categorial parser based on such lexical entries obtains **good results** in:
  - question answering,
  - textual entailment.

Example 3

Paperno, Pham, Baroni 2014 (ACL 2014, next week):

- title: "A practical and linguistically-motivated approach to **compositional distributional semantics**",
- nouns are co-occurrence vectors (standard DS),
- adjectives (semantically take nouns as arguments) are matrices (Baroni Zamparelli 2010),
- so that adjective–noun combinations are vectors (results of matrix by vector multiplication),
- scale problem:
  - if nouns have 300 dimensions (vectors of length 300),
  - then adjectives have 300^2 dimensions,
  - transitive verbs (take 2 arguments) – 300^3 dimensions,
  - adverbs modifying transitive verbs – $(300^3)^2$ dimensions, etc.

Example 3 (contd.)

Paperno, Pham, Baroni 2014 (contd.):

- this paper:
  - a method of **scaling up** compositional distributional semantics,
  - based on classical **compositional semantics ideas**.
- **pros and cons**:
  - compositional semantics **without manually constructed grammatical resources**,
  - so far applied to **relatively simple tasks** (sentence similarity, basic entailment) – not clear how it extends to full textual entailment, question answering, etc.

Example 4

Raymond Mooney at al., since 2006:

- 2006: "**Learning Language from Perceptual Context**: A Challenge Problem for AI" (AAAI Fellows Symposium),
- idea: "learn language like a human child", from utterances paired with perceptual context,
- aka **grounded language learning**;
- includes work on **learning semantic parsers** with ambiguous semantic representations, e.g.:
  - RoboCup games wih **commentaries**,
  - each commentary is **paired with semantic representations** of actions happening within the last 5 seconds,
  - one of these representations usually corresponds to the commentary (**ambiguity, noise**).

Currently:

- very **limited domains and tasks**,
- but **much interesting research** at various places.

Conclusion

Theses of this talk – **in order to advance HLT, we need**:

- more emphasis on **language-specific resources and methods**, and on ways of **combining** them with "language-independent" **statistical approaches**,
- more emphasis on **compositional semantics**, and on ways of **combining** (or obtaining) it with statistical and **machine learning** approaches (also including **perceptual input**).

Thank you for your attention!