

Many Faces of Text Processing

Marko Grobelnik
Artificial Intelligence Lab

Jozef Stefan Institute

Ljubljana, Slovenia

+38614773900

marko.grobelnik@ijs.si

ABSTRACT

Why people process text with computers? It all started many years ago, with the main goal in minds of researchers, to understand the text. In the meantime, the area of text processing developed in many different directions whereby the original goals were often forgotten. Funny enough, it seems, in several decades of computerized processing of textual data, the solution to the 'text understanding' problem didn't evolve much compared to some other, easier and often more profitable problems to deal with (such as information retrieval/search, machine translation or information extraction). In this paper we touch various aspects of text processing along several dimensions: (a) how we represent the textual data, (b) what kind of algorithms and techniques we use, and (c) what kind of problems we solve on the top of text. Finally, it is interesting to observe various research communities dealing with textual data in different ways. Most of them are still rather fragmented and don't learn enough from each other - many of the ideas developed within one community don't cross borders of that community for too long.

Categories and Subject Descriptors

I.2 ARTIFICIAL INTELLIGENCE, I.2.4 Knowledge Representation Formalisms and Methods, I.2.6 Learning, I.2.7 Natural Language Processing,

H.3 INFORMATION STORAGE AND RETRIEVAL, H.3.1 Content Analysis and Indexing, H.3.3 Information Search and Retrieval

General Terms

Algorithms, Human Factors, Languages.

Keywords

Natural Language Processing, Text Understanding, Text Representation, Text Processing Algorithms, Text Processing Research Communities

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS'11, May 25-27, 2011 Sogndal, Norway

Copyright © 2011 ACM 978-1-4503-0148-0/11/05... \$10.00.

1. INTRODUCTION

Automatic text understanding has been a research goal for many years and it seems that it will continue to be explored for many years to come. The problem is hard because of two main issues: (a) the ungraspable nature of the world we are living in, and (b) the dynamic nature of human languages, trying to capture abstractions about the world for the purpose of communication between humans.

On this long path of exploring language and simultaneously modeling the world, we have transitioned through many stages. From the early years, when linguists were exploring the structure of words and sentences, and philosophers were discussing ontology, up till today, when a large part of written information is available to almost everybody and modern communication allows researchers to join their efforts and exploit opportunities offered by technology. Still, the apparent opportunities seem to be much bigger than the actual depth of the solutions developed in recent years in the field of text understanding.

There are three main dimensions along which we classify approaches to text processing:

- **Representations** (how we internally represent text?): characters, words, phrases, bag-of words, part of speech, parse trees, subject-predicate-object triples, semantically annotated triples, event frames, first order theories, complex events
- **Techniques** (how we manipulate text representations?): crowd-sourcing, analytic approaches, part-of-speech tagging, deep parsing, semantic annotation, semantic disambiguation, event extraction, reasoning, visualization
- **Tasks** (what problems we try to solve?): document search, clustering, document labeling, categorization, visualization, summarization, translation, categorization, contextualization, personalization, trend detection and monitoring, news bias detection, plagiarism

Another important dimension for understanding origins of the above dimensions is more of a social nature - it is about research communities addressing different parts of the text related domain. More on this will be discussed in the next section.

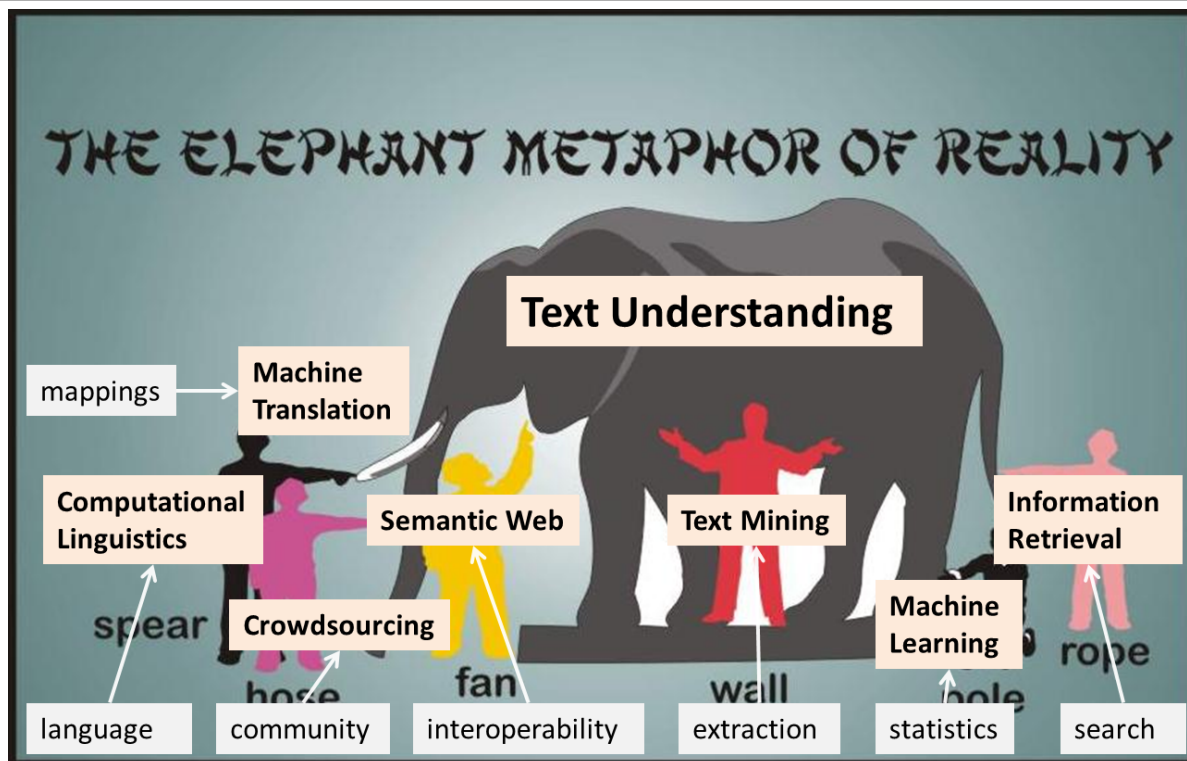


Figure 1: Using metaphor from the story about “Blind men and an elephant” to show partiality of the research areas dealing with text

2. Research Communities dealing with Text

Having many techniques for dealing with text, has the unfortunate consequence that those techniques fall into different research areas, which often do not communicate effectively. Text understanding, as a long-term research goal, can significantly benefit from a diversity of insights coming from related research areas.

So we can say, using a well-known metaphor from the story about “Blind men and an elephant”¹, that research areas dealing with text address only parts of the problem (the same as blind man in the story) and interdisciplinary cooperation among them would greatly contribute to the solution of the whole problem. Figure 1 shows the main research areas dealing with textual data with key elements used to build theories upon. Each community approaches the problem of dealing with text with different techniques, representations and tasks to solve.

Some of the examples of the above mentioned research areas with its characteristics are: 'Information Retrieval' field addresses many aspects of searching document collections while typically using rather simple representations (augmented bag-of-words stored in inverted indices), while 'Computational Linguistics' would often use rich representations on smaller units of text (e.g. sentence level), or 'Semantic Web' which uses rich symbolic representations mostly for the task of interoperability and reasoning.

All of the mentioned (Figure 1) research areas generally fall under one of the following three key scientific paradigms:

- **Top-down approaches:** Semantic Web, Knowledge Representation & Reasoning, Traditional Natural Language Processing
- **Bottom-up approaches:** Machine Learning, Data Mining, Information Retrieval, Modern Natural Language Processing
- **Collaborative approaches:** Web2.0, Social Computing, Human Computation

In the future we can expect integration among the above three different paradigms which will bring additional insights into how to solve original task on 'text understanding'.

3. Representing text

One of the key properties of natural languages is redundancy in the encoded information and the structure used to connect different parts of information (often with inherent exponential statistical distributions). As a consequence, number of different techniques can be applied to extract different types of information from a text. These techniques range from simple ones such as character counting, to more sophisticated ones such as linear algebra, to the advanced techniques which exploit the structural aspects of text.

Many of these techniques deliver something useful and solve somebody's problem. Some examples of such problems are: language identification (solved with simple character counting), document categorization (solved with linear algebra methods), question-answering (solved typically with

¹ http://en.wikipedia.org/wiki/Blind_men_and_an_elephant

- ▶ Characters (character n-grams and sequences)
- ▶ Words (stop-words, stemming, lemmatization)
- ▶ Phrases (word n-grams, proximity features)
- ▶ Taxonomies / thesauri

Lexical

-
- ▶ Part-of-speech tags
 - ▶ Vector-space model
 - ▶ Language models
 - ▶ Full-parsing
 - ▶ Cross-modal processing

Syntactic

-
- ▶ Collaborative tagging / Web2.0
 - ▶ Templates / Frames
 - ▶ Ontologies / First order theories

Semantic

Figure 2: Different levels of representing text

shallow linguistic methods), and reasoning (solved typically using logic).

Figure 2 shows the stack of text representations sorted from simple ones (character based) to more sophisticated ones (ontologies and first order theories). The representations listed on the figure 2 are clustered into three major groups according to complexity of processing attached to it:

1. Lexical based representations where the unit of processing is a character, a word or a phrase (sequence of words). The key characteristic of lexical representations is to deal with text as a set of data items, possibly extended with simple preprocessing (such as dictionary look-up, phrase generation etc.)
2. Syntax based representations take into account shallow, surface based structure of the text. The most obvious and popular characteristic is word co-location which is extensively used in information retrieval and machine learning. More sophisticated representations include language specifics (such as PoS tagging and parsing).
3. Semantic based representations go beyond surface representations and represent semantics of text in various forms. The most simple type of semantics is represented through word relationships, more sophisticated include semantic frames (as collections of related information slots describing particular phenomena) or logic based ontologies or theories providing deeper or shallower background knowledge on how fragments from text (either words, sentences or documents) are inter-related.

All of the above representations map to typical techniques which further allow solving particular tasks.

4. Conclusion

In this short paper we touched the breadth of the area of text processing. It spans across several research communities, it includes several representations and even more techniques. The tasks which are being solved nowadays range from very popular ones such as web search (used by most computer users) to very research based (such as reasoning with text represented in first order logic).

In the future we expect integration of ideas and paradigms from several research areas which will us lead closer to parts of a solution of the problem of 'text understanding'.