Spring
2017

Natural Language Processing

ADVANCED TOPICS IN ARTIFICIAL INTELLIGENCE

ELEFTHERIOS TRIVIZAKIS, ΜΠ143

Content Table

A brief history

Natural language processing started in 1950s as a domain from the combination of artificial intelligence and linguistics. Originally, it was separated to textual information retrieval but over time those two fields converged. Current NLP borrows methodologies and techniques from diverse fields [1].

Theoretical analysis of language grammars like Chomsky's in 1956 provided an estimation of the difficulty of the problem. Also, it influenced the creation of Backus-Naur Form (BNF) notation (1963) to specify "Context-Free Grammar" (CFG), common for representing programming language syntax [1].

The BNF specification is a set of derivation rules that collectively validate program code syntactically. Also, Chomsky identified a more restricted "regular" grammar used to specify text search patterns.

During the 1970s lexical-analyzer generators was invented which transforms a text into tokens and parser generators which validates a token sequence. Prolog [8] was created as a natural language project with syntaxes especially suited for writing grammars.

In 1980s simple approximations replaced deep analysis, evaluation became strict, creation of machine learning (ML) methods utilizing probabilities, large annotated bodies of text (corpus) became available for training ML-algorithms, built the foundations of statistical NLP which is strong if large corpus is available but not suited for unknown inputs. Most important became wide spread the realization of handwritten rules and statistical approaches are complementary to each other [1].

An introduction to natural language processing

Natural Language Processing is developed for the purpose of learning, understanding and processing unstructured human language content [2].

The first applications targeted automating the analysis of the linguistic structure of language and building services like machine translation, speech recognition and speech synthesis. The modern research community utilizing those tools in real-world applications creating spoken dialogue systems and speech-to-speech translation engines, mining social media for information, and identifying sentiment and emotion toward products and services [2].

Computational linguistics have grown into both scientific and practical technology embedded into consumer products. Four key factors are responsible for these developments increase in computing power, availability of large amounts of linguistic data, the development of efficient machine learning techniques and a better understanding of structure of human language and its advances in social context [2].

The ultimate goal for NLP is to aid the human to human and human to machine communication and benefit both by analyzing and learning from the vast quantity of human language content that is widely available through multiple sources and online databases [2].

Natural language vast size, unrestricted nature and ambiguity led to two problems when using standard parsing approaches that rely purely on symbolic rules: meaning extraction and rules that can't handle "ungrammatical" spoken prose [1].

The major drawbacks for developing NLP services and applications are highly related to variability, ambiguity and content-dependent interpretation of human languages and human-made content. Other challenges are finding a representative corpus (bag of words), understanding the linguistic structure, identify syntactic and semantic information, identify the context [2].

To encounter all those obstacles there is a need for incorporating the large amounts of digital texts and transform them as linguistically annotated data. Utilize large speech and text corpora with Part-of-Speech tagging, syntactic parses, semantic labels, annotations of named entities like persons, places, objects, dialogue acts, emotions, positive and negative sentiment and discourse structure [2].

Natural language processing tasks

There are four main techniques applied to NLP application tokenization, word segmentation, part-of-speech (POS) tagging and parsing [3].

Tokenization [3] is a fundamental technique for NLP tasks, it splits a sentence or document into tokens which are words or phrases. For English splitting words by spaces is trivial but additional knowledge should be taken in consideration like for example opinion phrases or named entities. Also, small words with no meaning contribution will be removed.

Word segmentation [3] is a sequential labeling problem with approaches like conditional random fields, hidden Markov, maxim-entropy Markov models, word embedding, deep learning for Asian languages.

Part-of-Speech tagging [3] along with parsing are techniques that analyze the lexicon and syntactic information. POS is used to determine a corresponding tag for each word. Similarly, to word segmentation is a sequential labeling problem. The tags include adjective, noun, verb and are helpful for example in opinion mining mostly because adjectives are opinion word and noun the targets.

Parsing [3] obtains syntactic information and produces a tree which describes the grammatical structure of any sentence with the corresponding relationship of different constituents. Comparing with POS tagging, parsing provides richer structure information.

Other low-level NLP tasks consists of sentence boundary detection, shallow parsing or chunking for identifying phrases and problem specific segmentation for creating meaningful groups of texts [1].

The secondary higher-level processing techniques includes spelling or grammatical error identification and recovery, named entities recognition (NER) which maps entities with word phrases, derivation from noun to adjective, inflection, synonymy, homographs of polysemy or abbreviations, word sense disambiguation, negation and uncertainty identification, relationship extraction temporal inferences and information extraction [1].

Namely, some toolkits for NLP implementations are PenNLP, CoreNLP, Gensim and FundanNLP written in Java, LTP for python and NiuParser for Chinese written in C++ [3].

Concluding, other data driven approaches are proposed like N-grams for suggest autocompleting and spelling, chaining NLP for pipelining, support vector machines for linear separation or Gaussian, hidden Markov models for inference, pattern matching, naïve Bayesian, conditional random fields as mentioned earlier [1].

Active learning techniques

The main problem with NLP applications is building an annotated lexicon or a sufficient meaning representation. Active learning try to bypass these limitations introducing the idea of a system trained with a small amount of annotated data utilizing classifier modules [4].

Following the training the system is tested with unannotated examples and attaches certainties to the predicated annotation of those examples. Finally, the examples with the lowest certainties are presented to the user for manual annotation and repeat the training [4].

An implementation of such a system in prolog is the *CHILL* parser, which given a set of training sentences each paired with a meaning representation maps those sentence into a semantic form. Inductive logic programming methods are used to learn a deterministic shift-reduce parser. The learning rules control the initial general parsing shell, facing challenges like the parser acquisition. Also, it’s important that user contribution must be provided [4].

A different pattern recognition approach is *Rapier* which utilizes a bottom-up relational learning and acquisition rules for information extraction in a natural language document or database. It obtains structured knowledge bases from unstructured pools of data [4].

Concluding, the main focus for active learning for natural language processing, namely, are information extraction, named entity recognition, text categorization, art-of-speech tagging, parsing, word disambiguation spoken language understanding, phone or other structured sequence recognition, automatic transliteration and sequence segmentation [5].

Applications

Machine translation [2] purpose is to aid human communication. Correct translation requires not only the ability to analyze and generate sentences in human languages but also a human-like understanding of knowledge and context despite the ambiguities. The evolution of these services from early implementations and word to word to today's phrase-based translations. The target of research is to build deeper meaning representations of language for a new generation of semantic machine translations.

Spoken dialogue systems and conversational agents [2] for text-based to spoken dialogue systems (SDS) will improve the human to machine interactions. This requires automatic speech recognition for determining what the human is trying to say, a dialogue management mechanism for understanding what the human wants, text-to-speech to communicate answers in a human familiar way and the ability to interact and respond with the user even partially information or errors occur during the interaction, in other words context awareness.

A serious constrain is the location-sensitive nature of these implementations which are more suitable for indoors use where the noise levels and the privacy are more or less given. Potential positive side effects could be the sense of companionship for people that are unlikely to socialize in other ways.

Basically, the problems that arise is recognizing and producing normal human conversation behaviors, turn taking and coordination, interpret subtle cues in speaker's voice, expressions and body language, responsiveness and disambiguate words with diverse meaning.

Machine reading [2] is the idea of integration and stigmatization information for humans by comprehensively understanding large quantities of text. Although there is a large structured Knowledge Base expressed in a formal logical language it is impossible to compare with the huge repositories in human languages. Researcher targeting to exploit AI for extracting information from databases and perform subsequent reasoning and hypothesis generation.

Crucial for machine reading [2] is the relation extraction task which has to be general and semantically precise if aiming to extract all relations from a piece of text, also refereed as open information extraction. It can be achieved by making use of linguistic structure and should fulfill the criteria of high scalable, fast extraction and working with unlabeled data.

Mining social media [2] is a modern application aiming services like facebook, twitter or youtube for extracting useful knowledge. Some research fields can identify demographic info, large quantities and classes of data, language specific content, track trending topics or popular sentiment, identify opinions and beliefs, predict event or disease outbreaks, recognize deception in fake reviews or news and identify social networks of people.

Some disadvantages concerning privacy issues, unauthorized control over personal data, bias due to availability of data by services, challenges related to discovery of the “grand truth”, no validation of posters identity or the validity of posts and different and multiple sources at any given time.

Analysis and generation of speaker state or private state [2] refers to opinions, speculations, beliefs, emotions and any other evaluative views that are personality held by the speaker or writer.

A lot of effort from NLP researchers is focused on sentiment analysis, the identification of positive or negative orientation of textual content and in identifying beliefs states like neutrality, committed or uncommitted belief, based on lexicon and syntactic information.

Sentiment and belief constitutes by attitudes towards events and propositions or concerns attitudes towards people, organization and abstract concepts.

The detection of those characteristics in text requires lexical and sentence-level information, can be signaled by words which convey positive or negative orientation. For this purpose, online sentiment dictionaries are developed to specify the sentiment or more sophisticated approaches for identifying the source or the target of sentiment.

Other characteristics that can be extracted are the Ekman's classic six basic emotions or universal emotions anger, disgust, fear, happiness, sadness and surprise, identify other speaker state of deception, medical conditions like autism, Parkinson's disease (mostly speech-to-text), age, gender, likability, pathology, personality, cognitive load, sleepiness, interest, trust.

Applications of sentiment classification [2] used in opinion identification include positive or negative views or institutions or ideas or products, predicting votes from parliament records or predicting supreme court decisions from proceedings, mining social media, public mood, predicting stock market trends, evaluating a community's mental state.

The difficult domain of semantics, context and knowledge requires new discoveries in linguistics and inference. Development of probabilistic approaches to language is not just about solving engineering problems but also problem in linguistic science.

Sentiment extraction using NLP [6] consists of a topic specific feature term extraction, sentiment extraction and a subject of sentiment provide by relationship analysis utilizing linguistic sources like sentiment lexicons or sentiment pattern databases. Mining from unstructured databases and free form texts is more cost and resources efficient in contrast with the traditional surveys.

There are two main challenges the identification of overall opinion against the individual aspects of topic and associating opinion to a specific topic-term in the same context among coexisting terms or contexts. The scope of sentiment analysis applies mainly to kernel sentences (one verb).

The resources needed for such application are the sentiment lexicon, sentiment pattern database, definition of the scope of analysis, sentiment phrases and relationship knowledge base acquisition.

Information retrieval [7] is the task of selecting documents or pieces of content from a database in response to a user’s query. The most approaches using statistical engines for efficiency because building lexicons or formal structured knowledge bases can be time consuming and expensive.

The implementation presented in [7] performs automated indexing of content then search and rank the user generated queries. Particularly, it uses fast parsing for grammar analysis for each sentence, word affix trimmer for extracting the root from each word, head-modifier structure that changes the sequence of verbs, nouns or adjectives, term correlations for contextual connection with terms and query expansion semantically analysis of user input.

Conclusion

As technology spreads in every aspect and form in our everyday lives, the unstructured user generated content will grow exponentially providing a vast pull of data with untap potential. The only way to exploit and transform all that noise into knowledge is by extending the understanding of humans themselves and their communication.

The summarization, semantic analysis, abstract meaning extraction and the unification of similar knowledge through wide and diverse sources are areas that machines potentially can thrive, leveraging the computational power, practically infinite memory storage and unbiased reasoning. Handling by machine, these repeatedly and time-consuming procedures, allow humans to focus at creating new knowledge and advances in each discipline.

Besides the current and upcoming difficulties, natural language processing [8] and understanding will be an important component of future computing, consumer products and personalized services. Even the modern personal computing includes some form of these applications either for providing useful information, executing complex interdomain calculation and decision making all accessible from a human-like machine to human and vice versa communication.

References

[1] P. M. Nadkarni, L. Ohno-Machado and W. W. Chapman, "Natural language processing: an introduction," *The Journal of the American Medical Informatics Association,* no. 18, pp. 544-551, 2011.

[2] J. Hirschberg and M. D. Christopher, "Advances in natural language processing," *Science Magazine,* pp. 261-266, 17 July 2015.

[3] S. Sun, L. Chen and C. Junyu, "A review of natural language processing techniques for opinion mining systems," *Elsevier,* vol. 36, pp. 10-25, 2017.

[4] C. A. Thompson, M. E. Califf and R. J. Mooney, "Active Learning for Natural Language Parsing and Information Extraction," in *ICML '99 Proceedings of the Sixteenth International Conference on Machine Learning,* Bled, 1999.

[5] F. Olsson, *A literature survey of active machine learning in the context of natural language processing,* SICS Report, 2009.

[6] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques," in *Third IEEE International Conference on Data Mining*, Florida, 2003.

[7] T. Strzalkowski and B. Vauthey, "Information Retrieval Using Robust Natural Language Processing," in *30th Annual Meeting of the Association for Computational Linguistics*, Deleware, 1992.

[8] B. Gamback, J. Karlgren, C. Samuelsson, “Natural-Language Interpretation in Prolog”, SICS - Swedish Institute of Computer Science