

In Defense of Symbolic NLP

Konstantin Bogatyrev

Universal Dialog, Inc., San Diego, California

The paper examines the benefits and the drawbacks of two competing approaches to natural language processing: statistical (probabilistic) and symbolic (deterministic). While the statistical approach is gaining popularity, better results may often be obtained using symbolic methodologies. The paper argues that the benefits of consistent deterministic parsers and lexica are well worth the time and effort required for their development. The Meaning \leftrightarrow Theory is specifically recommended as the best theoretic framework for cross-lingual NLP applications including machine translation and text retrieval. The paper concludes that best results are obtained using a combination of the two approaches when statistical methods are applied to the output of a deterministic parser.

Keywords: computational linguistics, machine translation, information retrieval.

Introduction

A shy couple in Tolstoy's "Anna Karenina" engages in a complex, linguistically sophisticated dialog by writing the initial letter of each word they do not dare to utter. A reader may view this episode as a romantic fantasy but, according to Tolstoy's biographers, this had actually happened to the author. Human ability to correctly identify one of several hundred words beginning with the same letter indicates a tremendous amount of redundancy built into natural languages. Unfortunately, most of this wealth of information eludes computer applications: natural language processing systems still struggle to identify correct meanings of fully spelled words in perfectly readable electronic documents.

So why did Tolstoy's characters (and their real-life prototypes) understand each other? It is widely assumed that the most obvious enabling factor in language communication is expectation: we only hear what we expect to hear and often fail to hear the unexpected. Insert a semantic non sequitur into an otherwise smooth conversation, and the listener will not be able to repeat it, even if the offending sentence is syntactically correct. Our ability to extract meanings that we expect to hear or read in a given context depends on what we have heard or read before. Therefore, it is natural to apply statistical data extracted from large amounts of existing textual information and use it to calculate the probability of a certain meaning in a specific context. This is how meaning extraction and disambiguation is handled by most NLP applications, and yet the performance of probabilistic NLP systems is often disappointing. On the other hand, traditional methodologies that rely on deterministic models of syntax and static lexical resources are often viewed as obsolete.

There are many explanations for the prevalence of statistical methods in natural language processing, including text retrieval (TR) and, more recently, machine translation (MT) compared to traditional (deterministic or symbolic) methods: lack of high precision scalable parsers and sufficient lexical resources; the necessity to maintain and manually modify large lexica and/or adapt them to a different subject area; complexity of required lingware development etc. Statistical systems, on the other hand, are highly flexible, language-independent and can be re-targeted to a new subject area or a new language in a matter of weeks. Additionally, there is no doubt that deterministic models of language

do not adequately represent the very important role that expectation plays in listener's ability to understand the speaker (which, as we have seen, is well known to novelists).

Still probabilistic methods alone will not be able to produce viable natural language parsers and in fact will not be able to successfully address any tasks beyond crude translations or low-precision search engines. This paper will attempt to demonstrate some areas where symbolic methods cannot be avoided. In fact, the only viable approach to high-precision NLP must rely on a combination of both approaches [1].

The following discussion will list some of the problems associated with purely probabilistic NLP.

The definition of event

Any formal model requires adequate representation of the objects it is supposed to handle. The success of a probabilistic model depends, among other things, on the way it defines *events*. A common type of event in probabilistic NLP is the occurrence of one or more words in a specific context. Here *word* usually means not a word-form (i.e. a character string as it appears in a text) but rather a stem or *lemma*. The context is in turn defined as an ordered set of adjacent words on both sides of each word occurrence and the length of such context window is usually set to an arbitrary chosen integer value. While equating a stem with a lexical unit (*lexeme*) may be a reasonable if not ideal way of representing meanings, the definition of *context* commonly used in statistical NLP is a clear oversimplification, which obscures important linguistic relationships. It is indeed reasonable to assume that the probability of a word having a specific sense depends on other words located in its proximity. However, this "bag of words" approach hides underlying complexities that require more precise discovery procedures. For example, it is possible that a probabilistic disambiguation algorithm will find that the meaning of "give" in the sentence

John gave a lecture on statistical NLP. (1)

is different from the same verb's meaning in

John gave Mary a ring. (2)

(cf. [2:239–241]) but the likelihood of the correct decision depends on some arbitrary extra-linguistic parameters, such as the length of the context window or size and quality of the training corpus. It is much less likely that the same algorithm will correctly disambiguate the meaning of "give" in

John gave Mary his lecture notes. (3)

By contrast, a dependency parser will easily determine that "give" in the example (3) has the same meaning as in (2) and therefore is not related to public speaking, since the word *lecture* in (3) is a dependent of the verb's object, not of the verb itself; moreover, this disambiguation choice will be made with a 100% certainty. On the other hand, only a probabilistic algorithm could detect the matrimonial flavor of the second example.

Semantic context and syntactic context

Existing statistical methodologies are reluctant to distinguish between *syntactic context* and *semantic context*. While the latter can only be studied through the analysis of co-occurrences, the former is often identified by deterministic procedures based on linguistic properties of sentence components. Obviously then, the accuracy of meaning extraction algorithms could improve if existing probabilistic methodologies were applied to the output of a deterministic procedure that identifies relevant syntactic components and their inter-relationships. However, in order to be effective, such procedure will require a much deeper level of syntactic analysis than stemming, part of speech tagging, and identification of certain syntactic relationships. For example, sense disambiguation systems that utilize some linguistic knowledge beyond part-of-speech tagging seem to perform not as well as the ones that completely ignore syntactic relationships [3].

Reasons why symbolic NLP is underused

There are in our opinion two fundamental reasons why deterministic methods are underused by the NLP community. First, for many scholars, deterministic syntax is synonymous with the generative theory. Unfortunately, Chomskyan linguistics with its rather arbitrary imposition of binary tree structures, a weak semantic component, and the lack of descriptive means for syntactic relationships is not an easy fit for the “real world” NLP.

Second, deterministic NLP depends in large part on the availability of a dictionary that provides both lexicographic definitions and some information about syntactic properties. Development of such dictionaries is a challenging task, requiring several person-years and a team of well-trained lexicographers. Few if any among the existing R&D teams have the means, the time, and the talent necessary for making this approach practical. On the other hand, attempts to use existing general-purpose dictionaries as lexical resources in symbolic NLP create a vicious circle of mutual dependency since their lexical definitions are written in a natural language and need themselves to be parsed in order to be useful for parsing of the target data.

Case in point: computational lexicography

Princeton WordNet, a computer-based semantic dictionary and the most popular lexical resource in the NLP community, is indeed an outstanding achievement in lexicography. However, WordNet was originally designed as a case study in psycholinguistics and the necessity to incorporate inflectional and syntactic information was realized by its creators when the project was well underway [3]¹. While lexical definitions in WordNet contain some syntactic data, such as sentence frames, it is not suitable for computationally efficient deterministic parsing. In other words, critically important syntactic information, such as the number and the type of verb’s nominal arguments cannot be easily extracted from lexical definitions.

Furthermore, none of the parsing models used in existing systems employs deterministic means of identifying collocations that are dependent on the syntactic context. For example, sense definition 7 in the WordNet entry for “school” states that it is only used to denote a plurality of fish and may not

¹ It is somewhat embarrassing, but perhaps not accidental, that the first computer thesaurus was conceived and created by psychologists rather than linguists. A lack of interest among the linguistic community in large-scale lexicographic work has significantly hampered progress in natural language processing. The ETAP linguistic processor described in the following sections is a rare example of a meaning-centered NLP relying on fundamental linguistic research.

appear in any other contexts. However that information is intended for a human English-speaking reader and WordNet does not provide an algorithmic procedure that would permit automatic extraction of such knowledge. Furthermore, the quoted meaning has nothing to do with the rest of the sense list. On the other hand, placing it under a different semantic category, together with other similar senses (i.e. *herd of deer*, *crowd of people*, *flock of birds*, etc.) requires descriptive means that could formulate contextual restrictions in terms of semantic and syntactic relationships, something that WordNet, or any other traditional dictionary for that matter, does not have. This example as well as many similar collocations, such as *take vacation*, *give lecture*, *have a lunch* represent a large category of collocations where one word determines the appearance of the other one in a certain syntactic position. Since even the best semantic dictionaries, such as WordNet, provide no consistent means of identifying these relationships, any sense identification algorithm based on traditional definitions will not be able to eliminate large amounts of noise. For example, an information retrieval system utilizing WordNet is likely to return documents about fishermen training in response to a query containing the phrase “school of fish”. Last but not least, none of the existing lexical resources is multilingual prompting developers of cross-lingual NLP to either align several monolingual dictionaries or to translate English resources, such as WordNet – with very limited success.

The Meaning↔Text theory

The Meaning↔Text theory (MTT) and an MTT-based universal multilingual linguistic processor known by the Russian acronym ETAP are very good candidates for a symbolic foundation of a combined NLP system due to their ability to successfully parse complex natural language structures using a dependency grammar and a dictionary which includes lexical descriptions and consistent syntactic information sufficient for deterministic parsing of most English sentences [5–9]. MTT views natural language as a converter between textual representation (phonetic, written, or digital) and semantic representation. Consequently, an MTT-base parser analyses texts on several layers with each intermediate layer operating on the output of the adjacent layers.

The layers may be roughly grouped into the following categories:

1. Morphology, which is a critically important link of the parsing chain that helps distinguish between semantic and syntactic components of a natural language text. Syntactically relevant information such as the type of nominal arguments depends on the output of a morphological component of a MTT-based parser. It is well known that inflected languages may generate hundreds of word-forms for a single stem and most statistical NLP applications, in particular in such areas as machine translation and cross-language information retrieval, usually rely on a deterministic stemmer. However, even in the English language known for its weak inflection, some types of sense identification can be performed without any reliance on the deep syntax layer. For example, the meaning of *define* in the sentence: *The peaks were sharply defined against the clear sky* is only compatible with the passive voice [10].
2. Syntax parsing is a complex task depending in large part on the quality of both syntactic rules and the dictionary. Syntactic rules are necessary to identify dependencies and to categorize their semantic properties. For example, syntactic patterns representing a verb phrase with the number and the categories of its argument dependents (the *government model*) will often suffice for identifying a specific sense. Most of the recent information extraction (IE) systems rely in shallow syntax parsing based on pattern matching. This approach proved useful in the development of effective IE algorithms for English but ran into problems with free word order languages where linear pattern definitions had to be extremely complex. Such problems are eliminated by the use of ETAP-generated syntax trees with lexemes (disambiguated dictionary entries) as nodes and syntactic relationships as branches.

3. Many of the important syntactic parameters defy any categorization and need to be treated as part of lexical definitions rather than syntactic rules. While traditional deterministic grammars have a poor track record in handling such lexico-syntactic anomalies, the concept of the *lexical function*, which is unique to the Meaning \Leftrightarrow Text model, provides a comprehensive formal framework for such phenomena. Lexical functions (LF) are collocations where the meaning of one word (*the value*) depends on the meaning and syntactic position of the other one (*the argument*). Lexical definitions in ETAP store functional values in lexicographic definitions of their arguments. For example, the definitions of the verbs below will *not* include the senses generated by the following LFs as they all happen to be uniquely defined by their arguments:

OPER₁ (attention) = pay

OPER₁ (claim) = stake

FUNC₀ (silence) = reign, etc...

Lexical functions, which, to the best of our knowledge, are not used in any other NLP system, provide a superior deterministic alternative to statistical treatment of collocations, as well as to semantic dictionaries such as WordNet.

Applying probabilistic techniques to the output of a deterministic parser

Existing probabilistic techniques, such as ranking methods used in text retrieval, operate on words or stems, while the output of deep parsing procedures, such as the one outlined above, consists of syntax trees with lexical entries as nodes. This may restrict usability of some popular probabilistic methods. For example, the hidden Markov model that is often used in cross-language information retrieval relies on estimates of translation probabilities from the target language(s) to the query language; this does not seem to be a good fit unless we assume that all cognates have the translation probability of 1. On the other hand, the *tf.idf* method that evaluates document relevance as a function of query terms' frequency in the document, is applicable for as long as syntactic structures used in matching can be associated with semantic concepts.

References

1. Konstantin Bogatyrev, Roman Yangarber: Use of deep syntax parsing in cross-language information extraction. *MLMTA 2005*: 18-24
2. Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.
3. SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems. <http://www.sle.sharp.co.uk/senseval2>.
4. Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. *Introduction to WordNet: An On-line Lexical Database* (Revised August 1993.) <http://www.cogsci.princeton.edu/~wn/papers/>
5. Mel'čuk, I. "The Meaning-Text Approach to the Study of Natural Language and Linguistic Functional Models. Invited lecture," *S. Embleton (ed.): LACUS Forum 24, Chapel Hill: LACUS*, 1998:3-20.
6. Apresjan Ju. D, I. M. Boguslavskij, L.L. Iomdin, A.V., Lazurskij, V.Z. Sannikov, and L. L. Tsinman. "The linguistics of a Machine Translation System," *Meta*, **37** (1), 1992:97-112.

7. Apresjan, Ju. D, L.L. Iomdin, A.V. Lazurskij, V.Z. Sannikov, et L. L. Tsinman. "Le système de traduction automatique ETAP," *La traductique. Étude et recherches de traduction par ordinateur*. Les presses de l'université de Montréal, 1993: 364–377.
8. Apresjan, Ju. D., I. M. Boguslavsky, L. L. Iomdin, A. V. Lazursky, L. G. Mitjushin, V. Z. Sannikov, L. L. Tsinman. "Automatic Dictionaries in the ETAP-3 System," *Papers from the 1993 Spring Symposium, Bonnie Dorr, Program Chair, Technical Report SS-93-02*. The AAAI Press, 1993:93–97.
9. Boguslavsky, I. M. "A bi-directional Russian-to-English machine translation system (ETAP-3)," *Proceedings of the Machine Translation Summit V*. Luxembourg, 1995.
10. An example offered by Ju. D. Apresjan and I. M. Boguslavskij (personal communication).