Journal of Intelligent & Fuzzy Systems 17 (2006) 575–581
IOS Press

Information structuring in natural language communication: Syntactical approach

Andrzej Bargiela[a] and Władysław Homenda[b,*]
[a]*Department of Computing and Mathematics, The Nottingham Trent University, Nottingham NG1 4BU, UK*
[b]*Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-661 Warsaw, Poland*

Abstract. This paper introduces a new framework for processing Natural Language statements. The parallel is drawn between the Natural Language processing and the Data Mining technology of information granulation. The formalism affords consistent representation of a well-known phenomenon of 'approximate' grammatical correctness of Natural Language statements. The ontology-based information structuring is a natural complement of the syntactical information granulation. The approach is validated on some simple Natural Language statements and the directions for the future development of the system are outlined. The paper focuses on conceptual framework only and, as such, is intended to stimulate further research into the various implementation considerations that are prerequisite of large-scale applications.

1. Introduction

Information Extraction (IE) from Natural Language (NL) statements is a process that happens very efficiently and almost subconsciously for humans yet it presents a formidable challenge for computers. This is because the use of the Natural Language constantly evolves and, while NL is constrained to some extent by the rules of grammar, it finds room for conveying the same concepts in many different ways. Conversely, the superficially similar NL statements may have very different meanings depending on the context and the modality of the specific utterance. So the challenge of IE is to process the natural language structures and to combine information that is found, explicitly stated or implied, into concepts that encapsulate derived knowledge [4].

One reason for which Information Extraction (IE) is of significant research interest is that it provides a basic reference for comparing different natural language processing technologies. However, a more fundamental reason is that IE focuses on the essence of intelligent information processing, that of formation of abstractions. In this sense IE parallels the endeavors of Data Mining that is primarily motivated by 'making sense of data'.

The rich track record of data mining research provides a valuable insight into the methodologies that lead to comprehensive and interpretable results and that ensure the transparency of final findings. In one way or another there arises an issue of casting the results as information granules – conceptual entities that capture the essence of the overall data set in a compact manner. It is worth stressing that information granules not only support conversion of clouds of detailed data into more tangible information granules but, very importantly, afford a vehicle of abstraction that allows to think of granules as different conceptual entities; see [1,10–14] and the references therein.

Clearly the task of information granulation is not a trivial one and it is dependent to a large extent on the application domain. Zadeh [13,14] promoted a notion of information granulation in the framework of fuzzy sets. Other formal and commonly exploited environments of information granulation deal with rough sets [9], set theory [5] and (interval analysis) [1,7,11]. In the context of granular computing the analysis of the Natural Language statements can be represented as operations on fuzzy sets. To make this point clearer we formalize the definition of a Natural Language. If a set of all

*Corresponding author. E-mail: homenda@mini.pw.edu.pl.

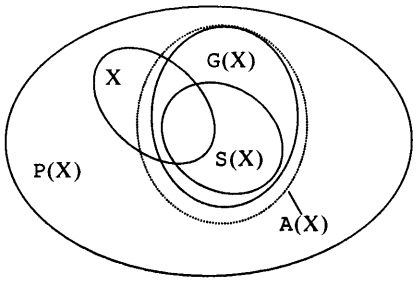

Fig. 1. Set representation of a natural language.

words in a given Natural Language is denoted by X, then the set of all possible utterances in this language is represented by a set of all subsets of X, i.e. the power set of X, denoted as P(X), as illustrated in Fig. 1.

Of course only a small proportion of elements of P(X) represent statements which conform to the rules of grammar of this language. These represent a subset of P(X), referred to as G(X). And the grammatically correct statements that are meaningful are represented by S(X) that is a subset of G(X). The words (X) themselves have a dual nature; on one hand they have grammatical meaning (parts of sentence) and on the other hand they have semantics defined by the concepts they represent. The latter is expressed in a thesaurus-style explanation through the related words. In this sense every word represents a fuzzy set defined over the whole domain X.

The essence of syntactical analysis is to discriminate whether a given sentence belongs to G(X) or P(X)-G(X). Syntactical analysis dissects the sentence into noun- and verb- phrases and checks for conformity with the rules of grammar until the bottom-most level of individual words is reached. A standard approach to syntactical analysis involves application of parsing techniques that use binary logic in checking the conformity of a given sentence with the rules of grammar. However, in real life many Natural Language statements can be considered meaningful even if they do not conform fully to the rules of NL grammar. So, in terms of the above set formalism, we suggest that the syntactical analysis should adopt a wider scope A(X) (of 'acceptable' statements) which is a superset of G(X), i.e. G(X) $\subset$ A(X). The degree of relaxation of the NL grammar rules is represented in terms of fuzzy logic operations where we no longer deal with binary predicates for a specific rule of grammar, but have a full spectrum of the 'degrees of truth'. One thesis of this paper is therefore that a fuzzy sets formalism applied in the context of parsing NL statements captures the natural tolerance to grammatical errors that one encounters in real life.

2. Natural language processing

Natural language is a tool supporting information representation and exchange in the process of human communication. On the other hand, natural language is rule driven, what is obvious in the context of its fundamental property of information exchange. Information encoded in a natural language construction (phrase, sentence, text) by a human being is addressed to other human being(s) with intention to be decoded and properly understood in their meaning. So, obviously, both subjects of information exchange supported by natural language constructions have to use the same rules in order to encode and decode respective information.

The goal of the natural language processing task is to design and build a computer system that will analyze, understand, and generate languages that humans use naturally. This goal is not easy to reach. "Understanding" language means, among other things, knowing what concepts a word or phrase stands for and knowing how to link those concepts together in a meaningful way. It is ironic that natural language, the symbol system that is easiest for humans to learn and use, is hardest for a computer to master. Long after machines have proven capable of inverting large matrices with speed and grace not achievable by human beings, they still fail to master the basics of human spoken and written languages cf. [3].

After decades of fruitful development of methods of natural language processing, it is clear that formalization of a full natural language and automation of its processing is far from complete. Natural language formalization and processing is often being restricted by different factors, for instance restricted to areas limited with regard to syntax, semantics, knowledge, style, etc., and even in these local areas of meaning, automation of its processing is still defective.

The set-theoretical representation of natural language utterances clearly leads to an enormous computational challenge. The generation and efficient handling of the power set build on the full vocabulary of the Natural Language is currently beyond the capability of current computers and consequently there is a large scope for research aiming at overcoming this computational challenge. Some obvious approaches, such as restriction of the original domain of discourse, have been successfully tried in the past but the inherent restriction of this approach proves too limiting in many situations. While recognising implementation considerations are an important issue, this paper focuses solely on the

conceptual aspects so as to stimulate the discussion of principles rather than implementation detail.

The challenges we face stem from the high flexibility and the ambiguous nature of natural language. Having English as his mother language, one effortlessly understand the sentence "Visiting aunts can be fun" assuming that you have some context knowledge about this sentence. Yet this sentence presents some difficulties to a software program that lacks both your knowledge of the world and human experience with linguistic structures. Is the more plausible interpretation that aunts are fun, or that rather visit is fun? Should the word "can" be analyzed as a verb or as a noun? Obviously, human being easily solves all these doubts with information recovered from contextual knowledge. However, information recovery that is subconscious for humans, raises challenge for automation, cf [5].

In our earlier publications [2] we have addressed also important aspects of "Social fundamentals" of natural language and that of "Lexical acquisition" and we assume that the reader is familiar with the works of [3, 6].

2.1. *Syntactical analysis*

Syntactical approach to natural language processing is the study of how words fit together to form structures up to the level of a sentence. Syntactical approach is a crucial stage and a crucial problem in natural language processing and in particular in extracting and representing information supported by natural language constructions.

The syntactical approach to natural language processing was extensively explored in the past until fairly recently, almost all work on automatic parsing has treated the task as essentially similar to 'compiling' programs in a formal computer language. Parsing was based on rules defining 'all and only' the valid grammatical structures in a language; faced with a particular input string, the task was to find the structural analysis by virtue of which it is a valid string.

2.2. *Context free grammars*

We focus attention on syntactical approach to natural language processing based on context free grammars CFG and transformations of CFG grammars. Of course, a grammar that is powerful enough to be able to analyse all English sentences is an impossibly large and complex, so we even will not try do construct it here. Rather we will try to develop a grammar that will meet the following three criteria:

- it will allow for analysis of all language phrases and sentences discussed in the paper,
- it could be developed to a more complete grammar that will extend the set of language phrases and sentences and will restrict generation of ungrammatical constructions
- it will use phrases and rules that are generally applicable in English, even if in some cases they involve a gross simplification of the way English works.

It is worth underlining that the grammar we develop generates a language that neither is included in English (the natural language), nor includes English. The language generated by the grammar intersects the English natural language and just their common part is a subject of our discussion.

Instead of a formal description of a grammar, which can be found in most introductory computer science texts, e.g. [6], we will be presenting only respective set of production or even only derivation of a language construction in the form of syntactical graphs. For extended description of context free grammar of English natural language cf. e.g. [3, pp. 51–53]. Note: it would be better to say – for context free grammar of a language that includes a subset of English natural language as its part. In light of previous comments, creating a grammar that fully describes English or other natural language is more sophisticated than using context free mechanism and more complicated than a few pages description. Anyway, this grammar will be referred as the grammar of English natural language, for simplicity.

3. Parsing natural language

Having a lexicon of words and a grammar describing a language, it is possible to formulate an algorithm to determine whether or not any given text is constructed according to the rules of the grammar. If a sentence is grammatical, the algorithm should be able to describe its structure. If a sentence is ambiguous, then the algorithm should be able to describe all its possible structures. An algorithm performing such a task is called a parser. As it was noticed, the lexicon and the grammar are essential elements of a parser and they decide about the quality of natural language processing.

3.1. *Syntactical structuring*

Syntactical structuring as a main task of parsing process has also its powerful contribution to information

extraction from natural language and – subsequently – to the process of information structuring and formation of a space of granular information. As it will be outlined below, syntactical structuring of language constructions strictly corresponds to granular structuring of information space. The following example gives intuition of the structure correspondence.

Let us consider the real sentence "The best students were operating the optical equipment". The basis of this sentence "Student operates" creates the simple sentence with relation between both words: noun *students* and verb *operate*. Then, the sentence is developed to build the more complex relation between noun phrase and verb phrase, each of them having complex structure. The central element of noun phrase – the noun *students* – is described by an adjective and a determinant. The central element of the verb phrase – the verb *operate* – is transformed to past –ing form and is supplemented by post-verb phrase. Despite of the complex structure of the sentence, the main relation is built on both central elements: noun *students* and verb *operate*. The additions wrap these central elements in extra information that define more specific meaning of – still the same – relation between noun and verb. And finally, the sentence has unique derivation in the grammar. Note: such features as adjective comparisons, noun plurality, numeral ordinality, etc. are omitted since they do not raise any novelty in the discussion.

3.2. Ambiguity

The trivial example "Student operates" can developed in order to specify included data. For instance, the sentence "The first student saw the man" develops embryonic sentence describing more exactly the noun *student*. And then the sentence: "The first student on the list saw the man with the camera" extends specificity of other parts of initial and subsequent sentences, cf. [3]. Armed with the grammar of English natural language, as e.g. in [3], and with a lexicon of English words and idioms, one can start parsing English natural language constructions. Parsing the last sentence leads to the syntactical graphs presented in Fig. 2:

Interpretation of this sentence is ambiguous: it is not clear whether the student was using camera when saw the men or rather the student saw the man and the man had camera. This ambiguity brings to two different syntactical graphs or – remembering that syntactical graphs are equivalent to derivations – to two derivations in context free grammar outlining the ambiguity.

The ambiguity could be resolved on the basis of contextual information. Considering both sentences "The best students were operating the optical equipment. The first student on the list saw the man with the camera." As a cohesive text, it would be easily deducted that the interpretation outlined in Fig. 2 is correct.

4. Information granulation

The naive sentence "*Student operates*" can be considered from the perspective of information supported by the sentences. On one hand, the pattern of language construction outlined in the form of syntactical graph defines relation between basic data of the noun type and basic data of the verb type. On the other hand, the sentence "*Student operates*" defines a relation between two simple pieces of data: "*student*" and "*operate*". Both pieces of data would be seen as elementary granules of data emerging from a plain surface of single words. In the first case, we have a pattern that defines a set of relations between words that can fill in the pattern. In the second case, the relation is defined on strictly defined words. We will focus our attention on the relation defined on words, phrases and sentences rather than on patterns of language constructions.

4.1. Granular space formation

It would be observed that natural language constructions provide the family of relations between words themselves creating tuplets of related words, between words and a tuplets of related constructions and between tuplets of constructions. It is clear that language constructions may have recursive structure, i.e. one phrase may include another phrase of the same type (noun phrase includes another noun phrase) or of different type (verb phrase includes noun phrase). And so, the structuring relations have a character of a tangled up net rather than simple hierarchical tree structure. However, a kind of a hierarchical structuring of these relations could be defined as elements of different language constructions supporting contextual knowledge. Considering the example of simple text of two sentences presented in Fig. 2, this hierarchy could be defined as:

- the set X of simple words, i.e. the set X = {be, best, camera, equipment, first, list, man, on, operate, optical, see, student, the, with},

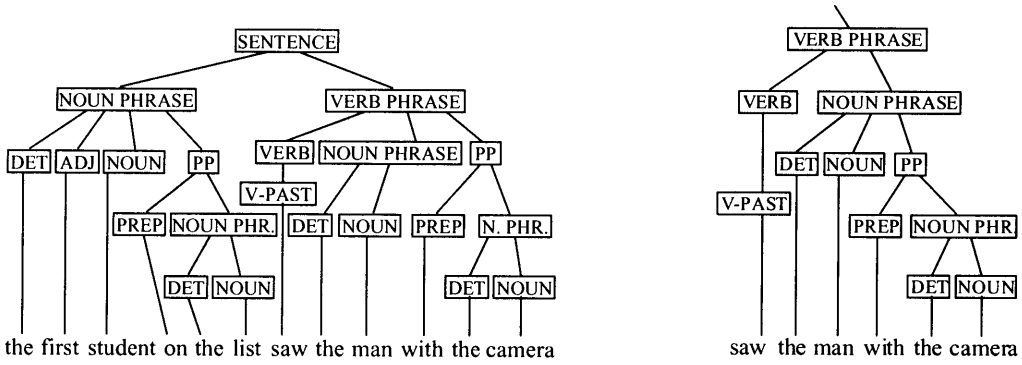

Fig. 2. Syntactical graphs of the sentence "The first student on the list saw the man with the camera". In the left graph the prepositional phrase "with the camera" is assumed to be a part of verb phrase "saw the man ...". The right graph presents the verb part of the syntactical graph: the prepositional phrase "with the camera" is assumed to be a part of noun phrase "the man ...".

- the set X with simple features added to form grammatical forms that can be seen as a syntactical graph defining proper grammatical form. In the presented example, these features can be illustrated by the following set of constructions: {operating, saw, the students, the equipment, the student, the list, the man, the camera, were},
- basic noun and verb phrases that do not include other noun or verb phrases as their parts. Again, they can be seen as syntactical graphs defining given phrases. The following phrases could be distinguished in the example: {on the list, the best students, the first student, the optical equipment, were operating, with the camera}
- compound phrases, i.e. noun and verb phrases that have other phrases as their parts. The example gives the following two compound phrases {the first student on the list, the man with the camera}
- sentences as pairs of noun and verb phrases. The example gives two sentences {The best students were operating optical equipment. The first student on the list saw the men with the camera.}

The granular space formed by language constructions is a kind of superstructure of the set of simple words and, from that point of view, can be considered as a dynamic extension of the static lexicon. This superstructure is built on the static fundament of lexicon every time new language construction is analyzed.

4.2. *Resolving ambiguity*

The ambiguity of the second sentence outlined in Section 3.2 can be easily resolved while both sentences are considered as a cohesive text. We operate with a kind of dynamical environment called "sentence neighborhood" that moves information between consecutive sentences. We assume that some piece of information, a granule, defined in a sentence, is valid in next sentences as long as it is not redefined. And then, if the first sentence defines the granule *the optical equipment*, this granule is still valid in the second sentence. Unlikely, the granule *the best students* is moved to the second sentence, but it is redefined to the granule *the first student on the list*. However, the redefined granule, as being more specific than its origin, inherits properties of its predecessor, so it still remains in relation with the granule *the optical equipment*. The lexicon dependency between camera and optical equipment allows for binding the prepositional phrase *with the camera* to *the student* granule rather than to *the man* granule.

It is worth underlining that newly defined granules create dynamic extension of the lexicon, as it was flagged in Section 2. And, of course, the newly added granules are bound with other granules of both static and dynamic parts.

These relations could be numerically described in the form of fuzzy sets having lexicon elements as their domains and labeled by given lexicon element. For instance, the granule *camera* interpreted as a fuzzy set may get the following membership values

$$\text{fs(camera)} = \{ \ldots, 0.9/\text{camera}, 0.7/\text{equipment}, 0.7/\text{optical}, 0.5/\text{student}, 0.5/\text{man}, \ldots \}$$

at the at the basic level of static lexicon. This fuzzy set can then be developed to

$$\text{fs(tcamera)} = \{\ldots, 0.9/\text{camera}, 0.7/\text{equipment}, 0.7/\text{optical}, 0.5/\text{student}, 0.5/\text{man}, \ldots, 0.9/\text{the camera}, 0.9/\text{the best students}, 0.9/\text{the optical equipment}, 0.5/\text{the man}, \ldots \}$$

when the first sentence is analysed.. The fuzzy sets represented the granule the camera will have similar numerical values of membership function. The second sentence includes its granules to the lexicon and, finally, will bind the granule *the camera* with other granules in the fuzzy sets possibly having the following membership functions:

fs(tcamera) = {..., 0.9/camera, 0.7/equipment, 0.7/optical, 0.5/student, 0.5/man, ..., 0.9/the camera, 0.9/the best students, 0.9/the optical equipment, 0.5/the man, ... 0.9/the first student on the list, ...}

Note: the numerical value expressing the relation between granules *the camera* and *the man* remains unchanged during all the process of text analysis and the dynamic extension of the lexicon. On the other hand, the sentences presented to text analyser increase the numerical value of the relation between wrapping granule based on the noun *student*. Thus, the ambiguity could be solved by simple comparison of membership values of certain fuzzy sets. The same solution could be reached when two fuzzy sets *the man* and *the first student on the list* are utilized. Membership values in the point the camera would also express tighter link between granules *the camera* and *the first student on the list* rather than between *the camera* and *the man*.

5. Ontology-based information structuring

Ontologies are created to formalize the relationship between the concepts represented by words. The formalization uses axioms, which are statements that are universally accepted as true. Ontologies can be viewed therefore as representing human knowledge in a computer readable form. They also help to introduce structure into large knowledge bases so as to achieve improved precision and relevance of knowledge base enquiries. A common complement of ontology is a lexicon. This provides mapping of concepts defined in the ontology onto Natural Language. The dynamically evolving lexicon described in Section 4.1 can be associated with ontology to provide an evolving interpretation of concepts while retaining the granular building blocks of the domain knowledge.

The ontology of the simulation and modeling domain is based on a general purpose ontology referred to in literature as Suggested Upper Merged Ontology (SUMO) [8]. This implies that our ontology is represented in Knowledge Interchange Format (KIF). A distinguishing feature of SUMO is that it is mapped onto a large general-purpose lexicon WordNet. This enables the user to refer to concepts in the SUMO ontology using Natural Language statements in English. The simulation and modeling domain ontology is created on top of SUMO. This is accomplished by a corresponding extension of the lexicon with special terms that are embedded in the ontology using special relations. The rationale for the above incremental extension of the SUMO is the intention to represent the knowledge about the simulation and modeling methodology and not just about narrowly specified mathematical models.

There are four major taxonomies constituting the simulation and modeling ontology. The first taxonomy is that of modeling paradigms: block diagrams, network graphs, differential equations, bond graphs, etc. Although it is possible to describe all the above elements quite formally we decided to provide only basic, intuitive description of concepts. The second taxonomy is that of typical modeling tasks and solution methods. The third taxonomy is that of components and processes referred to in second taxonomy. The fourth taxonomy is that of structured metadata associated with individual documents in the knowledge base.

The ontology-based information structuring outlined here is a natural complement of the syntactical information granulation. As the information granules evolve dynamically through the process of syntactical analysis of Natural Language statements, it becomes important to relate the concepts, represented by the granules, to the ontology representing the understanding of the knowledge. Using the set representation of Natural Language illustrated in Fig. 1, the ontology-based structuring helps to delineate the fuzzy boundary between S(X) and P(X)-S(X).

6. Conclusions

The paper casts the natural language processing problem in a novel framework of fuzzy sets and fuzzy logic based information processing. Simple examples considered in the paper indicate the feasibility of the task but a further research is needed to investigate the methodology for building the lexicon of fuzzy sets. Also an investigation into fuzzy parsing, that will capture the varying degree of tolerance to grammatical inconsistencies, will need to be undertaken. The combined conceptual and implementation challenge of the proposed framework is deliberately dissected into two constituent parts and the paper has focused on the dis-

cussion of the conceptual aspects. Discussion of implementation challenges and alternative solutions will be a subject of a separate publication.

Acknowledgment

Support from Natural Sciences and Engineering Research Council (NSERC) and Alberta Software Engineering Research Consortium (ASERC) is gratefully acknowledged.

References

[1] A. Bargiela, Interval and Ellipsoidal Uncertainty Models, in: *Granular Computing,* W. Pedrycz, ed., Springer Verlag, 2001.

[2] A. Bargiela and W. Homenda, *Information structuring in natural language communication: syntactical approach,* Proc. of the International Conference on Fuzzy Systems and Knowledge Discovery, Singapore, November 18–22, 2002.

[3] C. Beardon, D. Lumsden and G. Ho;mes, *Natural Language and Computational Linguistics, an introduction,* New York, 1991.

[4] D. Bikel, R. Schwartz and R. Weischedel, An algorithm that learns what's in a name, *Machine Learning – Special Issue on NL Learning* **34** (1999), 1–3.

[5] W. Homenda, Databases with Alternative Information, *IEEE Transactions on Knowledge and Data Engineering* **3**(3) (September 1991).

[6] J.E. Hopcroft, R. Motwani and J.D. Ullman, *Introduction to automata theory, languages and computationm,* Addison-Wesley, Boston, 2001.

[7] R. Moore, *Interval Analysis,* Prentice Hall, Englewood Cllifs, NJ, 1966.

[8] I. Niles and A. Pease, *Toward a standard upper ontology,* Proc. 2nd Int. Conf. on Formal Ontology in Information Systems (FOIS 2001).

[9] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data,* Kluwer Academic, Dordrecht, 1991.

[10] W. Pedrycz, *Computational Intelligence: An Introduction,* CRC Press, Boca Raton, FL, 1997.

[11] W. Pedrycz and A. Bargiela, Granular clustering: A granular signature of data, *IEEE Trans. Syst. Man And Cybern.* **B32**(2) (2002), 212–224.

[12] W. Pedrycz, M.H. Smith and A. Bargiela, *A granular signature of data,* Proc. 19th Int. (IEEE) Conf. NAFIPS'2000, Atlanta, July 2000, 69–73.

[13] L.A. Zadeh, Fuzzy logic = Computing with words, *IEEE Trans. on Fuzzy Systems* **4**(2) (1996), 103–111.

[14] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems* **90** (1997), 111–117.