Triangle: Language, Literature, Computation, n. 1, 2010
Publicacions Universitat Rovira i Virgili · ISSN: 2013-939X
https://revistes.urv.cat/index.php/triangle

An Introduction to Natural Language Processing: the Main Problems

Veronica Dahl

School of CS, Simon Fraser University &
GRLMC-Universitat Rovira i Virgili
veronica@cs.sfu.ca

1 Definition, Scope

Natural Language Processing aims to give computers the power to automatically process human language sentences, mostly in written text form but also spoken, for various purposes.

This sub-discipline of AI (Artificial Intelligence) is also known as Natural Language Understanding. So first of all, what do we mean by "understanding"? According to Longman's Dictionary of Contemporary English (1990), to understand is "to know or recognize the meaning of (something) or the words spoken by (someone)."

This definition plunges us into the ongoing debate about whether computers can know, think, feel, etc. We shall not attempt here to elucidate these deep questions. Instead, we will take the modest approach of acknowledging that no human or formal science has yet come up with a complete, satisfactory explanation of such human activities as thinking, knowing, and understanding.

However, we can still arrive at a useful working definition of what "understanding language" means for a computer, by noting that given a language stimulus, certain computer programs can replicate to some extent

some of the behavior that would typically be elicited by a human under the same language stimulus. Thus we can consider a computer program as being able to, in some sense, "understand" language, if its output given linguistic input roughly corresponds to a response that might be produced by a human given the same input.

For instance, consider a database query such as:

Do you know what time it is?

and consider the following possible answers:

1. *Five past eleven.*
2. *Yes.*
3. *True.*
4. *1.*

Assuming the time of the question is 11:05, the first is the kind of answer a human might produce, whereas the three remaining answers represent truthful but unhelpful versions of a literal answer: 2) replies in the affirmative, 3) gives a truth value representing the affirmative answer, 4) gives the same information in binary notation. We can think of these successive answers as diminishing in the degree of "understanding" shown, as measured by how close to human reaction the machine's reaction is.

Similarly, if we ask a machine to translate:

'La voiture n'a plus d'essence.'

we can conceivably get machine translations such as:

1. *The car has no more gas.*
2. *The car has no more essence.*
3. *The car has no more soul.*
4. *The car not has more of essence/soul.*

Here again, the degree of "understanding" evidenced by the translation given decreases as we proceed to the next example: "*essence*" is not the right word for translating '*gas*', but is a better match than '*soul*', whereas example 4) is simply a literal, word-by-word translation which does no justice to the original meaning.

This view of understanding language parallels the definition of intelligence proposed by Turing: if a machine can fool a human into believing he

is interacting with another human via a computer terminal, then it can be said to be intelligent.

But both Turing's test and the above proposed view of machine understanding of language have limitations. For instance, Weizembaum's famous Eliza program, which modelled a psychological counselor, could certainly fool a human into believing s/he was interacting with another human via a computer terminal, yet it only did this through parrot-like, key-word oriented sentence transformations which could by no means be called "intelligent". For instance, a sentence of the form: '*I feel X*', was simply transformed into '*Why do you feel X*', with no regard whatsoever to what it might mean, through a purely word-processing transformation.

More modern language processing systems are still concerned with human-like behavior, but also take inspiration from the results that other disciplines relating to language and cognition have to offer. For instance, we can take a linguistic theory, and try to tailor it to our own computational needs, under the assumption that this theory comprises a usable if imperfect model of what might be actually happening in a human brain.

We are now ready to define the discipline and its scope.

Definition

Natural Language Processing is a branch of Artificial Intelligence in which computers are programmed to simulate an "understanding" of human languages such as Catalan, English etcetera, for various purposes, such as consulting databases through human languages rather than some specialized computer interface, or translating text automatically from one language to another, or communicating across virtual worlds.

Scope

The areas in which NLP has been most successful are text processing (vs. spoken language), for isolated sentences (vs. dialogue), and with restrictions regarding both the subset of language addressed and the problem domain. The results are, however, interesting within the domains addressed, and the language is restricted usually in ways which still result in allowing input one would naturally write, as opposed to condensed or telegraphic versions. For example the passive voice and a number of other linguistic constructs, yet those allowed do permit to express all concepts needed in a natural fashion.

Our notion of machine "understanding", thus, while not being confined to behavior alone, is much more modest than that of human understanding: , it boils down to being able to respond to some extent as humans do, using adaptations of models of language provided or inspired by cognitive sciences, in particular linguistics. As pointed out in [5], "if machines are to communicate in human terms, they must embrace all the facets of natural language, and thus all of the parts of the broad topic of linguistics".

We shall next discuss some of these facets, in order to understand some of the reasons why the mechanical understanding and translation of human languages is not as easy as one might expect upon first examination. In this process we shall see that, for good reason, most problems that are central to AI are also central to NLP, which makes NLP doubly difficult: it must embrace all parts of the broad topic of linguistics, as well as those of the broad topic of AI.

2 The most common problem in NLP

The most studied problem in natural language processing is the parsing problem: given a grammar and a presumed sentence in the language defined by that grammar, obtain some representative structure(s) if the sentence is indeed in the language. Whether for parsing or other NLP problems, such as generation, translation, concept extraction, etc, we need to capture an infinite number of sentences with a finite device such as a grammar. This implies the need for a concise, regularity-capturing description means, such as can be provided by logic programming.

Why is it so difficult? Concrete examples

Communications in natural language tend to assume vast contextual and empirical world knowledge that can be taken for granted in a human, but must be somehow spelled out for a machine. We shall examine a few cases in which this spelling out can be difficult.

Ambiguity:

In the first place, human language is plagued with ambiguities that we are not always aware of, owing to the fact that the knowledge of the world that

we have and unconsciously use often prevents us from even seeing alternative possible meanings that a computer program would have to consider.

Take for instance the database query:

Which is the price of a cabinet with four drawers?

While most humans will only see one meaning in this apparently straightforward question, a machine will have to decide whether '*with four drawers*' modifies '*price*' or '*cabinet*'. That is, are we asking for the price of a cabinet such that that price has four drawers, or for the price of a cabinet such that the cabinet has four drawers? Obviously (to us), the first meaning is nonsense, since prices cannot possess drawers. But for a computer, unless we have somehow programmed into it the world knowledge that protects us from even considering the nonsensical sense, both meanings are, a priori, equally likely.

A poignant example of structural ambiguity is the case of propositional phrases, where ambiguity is combinatorially explosive. Take for example the following sentence:

I saw a man in the park with a telescope.[1]

This phrase can be interpreted in different ways according to which prepositional phrase attaches to which antecedent: either it tells us that the person was in the park and there saw a man who had a telescope, or that the person saw a man in the park and this park had a telescope, or that the person saw a man in the park through a telescope.

Different levels within Natural Language

We have no problem in simultaneously and effortlessly capturing the various levels speech involves (such as phonology, syntax, semantics, pragmatics, etc.), whereas precisely conveying them to a computer is quite difficult. In practice, most of the processing out there is syntactic, with some semantics to guide it. Speech recognition has lately advanced significantly, but still not enough to replace humans. The different levels can be characterized as:

Prosody: studies rhythm and intonation.

[1] The example belongs to Bill Wood

Phonology: studies sounds.

Morphology: studies syntactic subunits in a word (unity: a morpheme).

Syntax: studies rules for combining words into phrases and sentences.

Semantics: studies meaning.

Pragmatics: studies ways in which to use language with respect to world knowledge.

Note that none of these levels is trivial; e.g. according to how punctuation signs are placed in the following poem, we can find in it four completely different readings (Spanish speakers: find them!)

TRES BELLAS QUE BELLAS SON [2]
Tres bellas que bellas son
me han exigido las tres
que diga de ellas cual es
la que ama mi corazón

si obedecer es razón
digo que amo a Soledad
no a Julia cuya bondad
persona humana no tiene
no aspira mi amor a Irene
que no es poca su beldad

Pragmatic Knowledge, Implicit Meanings:

Another difficulty with processing language, which we have hinted at with our example: '*Do you know what time it is?*' concerns the need for pragmatic knowledge of the world to be shared with a machine. This knowledge is largely implicit and even unconscious in humans, so it is not easy to think of all its possible instances ahead and spell them out.

[2] Cited by Roberto Vilches Acuña in: *Curiosidades literarias y malabarismos de la lengua*, Editorial Nascimiento, Santiago de Chile, 1955.

A more radical example is the indirect request '*How cold it's getting*', uttered in the hope that the hearer will in response close an open window he or she is near to. Whereas in the time request example there is at least a mention to the information wanted (what the time is), in this case, the utterance contains no hint whatsoever of what is actually being asked.

As a final example, consider:

I was caught running a red light and the pig fined me for it.

Most people would not have any difficulties understanding who '*the pig*' is, and that this is not actually about running but rather about driving a car past a red light. All this would be very hard for a computer to infer.

Imprecision

While natural languages are inherently imprecise, most methods for processing language are unable to properly deal with imprecision. Statistical parsing approaches view precision as a measure of how good a parsing result is and usually do not concern themselves with how to more precisely, convey the meaning of an imprecise expression.

Zadeh [8] defines the notion of precisiation, within a theory called CW -Computing with Words-, as the conversion of a semantic entity, such as a question, proposition, command, etc. into another semantic entity which is computation-ready, that is, can be computed with.

For sentences that can be precisely interpreted, the notion of precisiation coincides with that of translating them into a semantic formalism which renders their interpretation and which can then be computed with by the usual means e.g. if the sentence is a query to a knowledge base, the evaluation of its semantic representation with respect to that knowledge base will compute into the answer to that query.

For sentences that cannot be precisely interpreted, existing machinery allows the precisiation of a simple imprecise sentence such as "Most Swedes are tall" such that it can serve as a basis for answering questions such as: "What is the average height of Swedes?" as concretely as possible, e.g. "Between approximately 170 cm and 200 cm" (also expressible by a user-friendly graphical interface provided by a specialized mouse called Z-mouse). However, this machinery is still being developed, and much needs to be done in terms of integrating it with contemporary NLU systems.

Coordinated sentences

Sentences with "and", "or", "but" and so on contain two or more sentences. They are a typical source of implicit meanings to be reconstructed, e.g. in: '*John ate a steak and drank a beer*', we know right away that **John** drank a beer, even though the subject in the second conjoint is left implicit.

Reconstructing the missing parts is not as easy as in the above example unless we have adequate pragmatic information. For instance, almost unconsciously we realize that '*cold hands and feet*' is shorthand for '*cold hands and cold feet*', whereas no implicit repetition of '*cold*' would even occur to us if we instead were to hear '*cold hands and fever*'.

Compound nouns

It is often difficult to see, in a sequence of several nouns, which of the nouns are head nouns and which are modifiers. For instance, Gerald Gazdar and Chris Mellish identify no less than 42 distinct structural descriptions in the innocent-looking phrase: '*Judiciary plea settlement account audit*' [3], with just binary noun compounding. Many of these would not even occur to a human but must be carefully sorted through by a machine.

Overgeneration

Another problem is that of avoiding overgeneration. During analysis this is often not crucial, assuming the sentences that are input are correct, but for synthesis we should not allow our grammar to produce more sentences than the correct ones.

Long distance dependencies

Finally, we must provide a means for recognizing relationships between parts of the sentence that may be arbitrarily far away from each other, e.g. in order to relate a pronoun with the noun phrase it refers to, or in order to allow topicalization, as in:

Logic, we love.
Logic, I know we love.
Logic, I suspected he knew we love.

where the direct object of '*love*' has been displaced for emphatic effect, and can appear at an arbitrary distance from it.

Stylistic resources

To complicate matters, stylistic resources such as metaphor, irony or allusion may also catapult a text beyond its literal meaning, thus making it difficult to know how much and which kind of background knowledge will be needed for a useful account of this meaning. Consider for instance the following excerpt from Juliet's Monologue, scene V, of Shakespeare's *Romeo and Juliet*:

> *Love's heralds should be thoughts, which ten times faster glide than the sun's beams, driving back shadows over lowering hills.*

Clearly such texts are beyond literal interpretation: thoughts are being personalized- they do not in fact glide, except metaphorically; sun's beams do not '*drive*' shadows back, nor do the hills '*lower*'. Topicalization (the movement of a phrase outside from its more habitual order, for the purpose of giving it emphasis) is also present: '*ten times faster*' has been moved from where it normally would be located and now precedes the verb '*glide*'.

As an example involving humor, consider the following notice, found at a public service office (and from which I should take inspiration for charging my students ☺):

> Answers: 1 euro
> Answers that require thought: 2 euros
> Correct answers: 4 euros
> Discussions: 40 euros
> Awkward smiles: free

Any human can immediately see the humoristic implications of this notice, but they would likely be lost to a computer.

Intention

Intention is paramount in human communication. Grice has uncovered principles of cooperative communication, and the usual assumption underlying NLP systems is that the intention is to communicate.

However, a message can be understood but stonewalled, as in the following dialogue between a new schoolteacher and one of the students in his class (example taken from the book *Le Petit Nicolas*):

> *Je m'appele M. Leblanc- et vous?* (My name is M. Leblanc- and yours?)
> *Nous non.* (Ours isn't.)

It is obvious that the student knows the request is to know what their names are, rather than to know whether they are called M. Leblanc as well. The question's misinterpretation is intentional, and for a machine to analyze the reply's real meaning, it would need to be given very subtle contextual knowledge of beliefs and intentions.

The Cultural Dimension

This human reliance on world knowledge becomes even more difficult to emulate by a machine when different views of the world intervene. In the time request example, the knowledge of its implicit meaning is fairly universal. Almost anyone can interpret the meta-message contained in that literal message.

But when different cultures or world views intervene, interpreting meta-messages is not such a simple matter. Sociological research by [6], for instance, analyses how some meta-messages typical of one sex tend not to be understood by members of the opposite sex. These misunderstandings are explained in terms of the thesis that communication between the sexes is essentially cross-cultural communication, given that boys and girls grow up in what are essentially different cultures (manifested for instance through different kinds of games), and are socialized in different ways. Consequently, one sex leans more towards a hierarchical and problem-solving interpretation of the world, while another leans towards an interpretation based on relationships and networking. These differential upbringings result in different world views and a consequent difficulty in interpreting each other's meta-messages.

Now, if members of the same social group can have trouble understanding each other's meta-messages simply because of gender differences in upbringing, what can we expect for more patently cross-cultural communication? Many of us have direct experience of the problems encountered in a foreign country, even when our own language is spoken there. For instance, a reply of '*Thanks*' to a simple offer for a second helping of food may well mean '*Thanks, yes*' in one dialect, while meaning '*Thanks, no*' in another dialect of the same language. And even within the same country, or even city, we encounter language differences that can be quite marked, as in cockney versus the prescriptive norm, also known as "BBC English", or as in juvenile jargons vs. adult talk.

Rigorously speaking, then, all these pragmatic, stylistic, gender, dialectal, social class, and other differences would have to be encoded in some

computationally efficient way in order for a machine to be able to respond as subtly as a human. But of course, in machines we are content with much less than what we expect from humans. Typically, we reduce the domain of application in such a way that human-like reaction is made much easier. For instance, if we reduce ourselves to a hospital environment for a database consultation application, the subset of language to be considered, as well as the possible stylistic and other variants, reduces considerably.

In the next section we shall examine how much simplification we have been content with in the past, and how much more ambitious we can get given the state-of-the-art and the progress in hardware technology.

From all these examples we can see why it is so difficult to teach a computer to properly understand unrestricted language. It is a process that involves practically all aspects of human experience: thoughts, actions, feelings, beliefs, knowledge, expectations, time, learning, reasoning, metaphors, humor, irony, etcetera. Therefore, the central problems in Natural Language Processing include many of the central problems in Artificial Intelligence: how to represent knowledge, problem solving, reasoning, non-monotonicity, belief revision, planning, learning, ... in addition to its own specific problems.

3 Divide and Reign

Our problem being as formidable as we hope to have impressed upon the reader, the best hope we have of solving it is through dividing to reign, in particular by scaling it down to solvable size, and attacking different problems separately. This generates new sub areas within NLP.

For instance, one of the most widely studied areas in language understanding is that of question-answering systems, which usually model a subset of language specific to a given target domain (e.g. a medical domain).

While not easily adaptable to other domains, these systems have met with reasonable success, owing to the fact that the context of discourse in these systems can be largely predicted. This is useful for instance in resolving ambiguities due to polysemy (multiple meanings for one word). In a query system for a financial domain, for example, we could describe only one meaning- the most likely one- of the word '*bank*' (as a financial institution), on the reasonable assumption that most users of this system will use the word in its financial, not its geographic (as in '*the bank of a river*') sense.

We can also reduce the range of natural language to be accepted (for instance, only sentences in the active, not passive voice, will be allowed).

A further simplifying factor concerns the limitation to single, isolated queries, which allows us a smaller context in which to look, for instance, for pronoun reference.

Machine translation is a sub-area that has had a more eventful history. Its initial period was characterized by the naive view that one-to-one correspondences existed between languages, and all we had to do was to encode them. But for instance, languages such as Lithuanian or Russian do not have any articles. The colors that are recognized in a rainbow vary from one culture to another. The one-to-one correspondence view soon collapsed under the evidence that many of the concepts and the syntactic and lexical constructions used to cover them are language or culture dependent. The initial over-enthusiasm resulting from the naive view gave way to an equally excessive pessimism, and funds became very scarce.

The three main approaches to machine translation are the direct one, the rule-based one, and the inter-lingua, with statistical approaches pitching in as well [4].

In the direct approach, heavily language-specific programs are designed to translate from one given language into another. Word-by-word replacement routines are complemented with ad hoc transformations performed after lexical substitution.

In the rule-based, also called transfer, approach, a source-language sentence is translated into a machine-readable form corresponding to the source language. This form is then mapped into a machine-readable form for the target language, and then the output is generated from it. The intermediate, machine-readable forms are dependent on the language considered. Therefore, mappings need to be constructed for each source-language/target language pair.

In the inter-lingua approach, the source-language sentence is mapped into a language-independent representation, from which the surface structure is systematically produced.

Just as natural language query systems do, successful machine translation systems also restrict their scope to very limited domains of discourse (e.g. weather reports), and to specific subsets of natural language (e.g. the relatively unambiguous declarative sentences found in technical documents). As well, they usually depend on interaction with human users

and/or post-processing by a human translator in order to correct the system's errors.

Machine translation systems also usually pay the price of extensive development and tuning to particular clients. To date, these systems are not yet widely used, both because of the limitations described above, and because of the time and expense required to develop them. So for a translation system we might relax our NLP definition even further and suggest that, even if its output is not what would be produced by a human, it will be acceptable if it helps the job of a human translator, e.g. by reducing his/her job to correcting the system's mistakes.

4 Concluding remarks

We have discussed some of the main difficulties in formalizing and automating the tasks of processing human language. As we have seen, some of them are even hard to imagine for humans who are used to effortlessly using and interpreting language, since so much of the knowledge used in so doing is unconscious and resorts to fuzzily defined world knowledge.

One implicit notion throughout our discussion is the desirability of using the results of cognitive sciences as well, and in particular of linguistic theory. Theoretical linguistics has indeed developed remarkable insights on some very complex linguistic phenomena, and is developing in directions which are more and more compatible with computational linguistic needs. However, it is also fair to observe that these theories do not have as their goal or method to provide immediate comprehensive descriptions of actual natural language. A natural language processing system that must process actual text (e.g. spontaneous speech) with a minimum of coverage and accuracy, must solve many problems for which linguistic theory does not have even in principle solutions. Moreover, when building actual language processing systems, many instances are found in which the analyses of linguistic theory are contradicted by the data.

The main challenge is, then, to adapt the general analyses and insights from linguistic theory into actual language processing systems, and to delicately interweave the many independently explored facets of language processing.

Acknowledgments

Support from the European Commission in the form of the author's Marie Curie Chair of Excellence, and from both the Universitat Rovira i Virgili and the Simon Fraser University, as well as from the Canadian National Sciences Research Council, is gratefully acknowledged.

References

1. Allen, J. (1994) *Natural Language Understanding.* Benjamin Cummings, Second Edition.
2. Covington, M. (1994) *Natural Language Processing for Prolog Programmers.* Prentice-Hall.
3. Gazdar, G. and Mellish, C. (1989) *Natural Language Processing in Prolog.* Addison-Wesley.
4. Koehn, Ph. (2010) *Statistical Machine Translation.* Cambridge University Press, 2010.
5. McEnery, C. L. (1992) *Computational Linguistics: A handbook and toolbox for natural language processing.* Sigma-Press. Wilmslow, U.K., 1992.
6. Tannen, D. (1990) *You just don't understand- women and men in conversation.* Ballantine Books, New York.
7. Jurafsky D. and Martin, J. (2009) *Speech and Language Processing- An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Second Edition. Prentice-Hall.
8. Zadeh, L. (2010) *Computing with Words- A Paradigm Shift.* UC Davis CS Colloquium, January 7 2010.