Seeking the Meme

The Memetics of Language Processing

By MIT Student

Human interfaces remain by far the most crude and restricted element of computer systems today. In an age when a digital device that fits in the palm of the hand can record video, play music, and connect two people from halfway across the globe, interaction between man and machine is still largely restricted to hitting buttons and moving mice. The reason for this rudimentary design is the basic difference in how people and computers think. So disparate are their methods that any communication must be reduced to a series of key presses or pointer movements. Of course, individuals with the necessary technical expertise are capable of understanding the languages of machines, but there is yet to be a computer able to comprehend the complexities and ambiguities of human speech. The development of such a machine is one of the ultimate goals of the discipline known as computational linguistics. Experts in that field apply their intellects to natural language processing, the attempt to break down human language into basic, unambiguous elements of the sort which a computer is capable of understanding. While the existence of such elements might at first seem unlikely, given the indefinite and contextual nature of natural languages, it is actually an issue which has already been addressed in a different area of study: Memetics. Memetics describes languages as large complexes of interdependent memes, and, as memes are by their atomic nature precise and unambiguous, such an approach clearly shows that the efforts of computational linguists are not in vain, that basic, exact units of language do really exist. Furthermore, meme-theory does not merely prove the existence of such units, but governs them; for the description of these basic language elements places them squarely within the definition of memes. Therefore, the practice of natural language processing is, at its core, an attempt to identify the constituent memes that make up human language.

Computers are devices of precision. At the most fundamental level they operate on the instructions of numbers. In the language of 0s and 1s there is no room for ambiguity or inexactitude. However, it is an inescapable fact that we humans, for whose benefit computers are designed and built, are neither inclined to nor particularly capable of the same perfect logic as they. Rather, "natural language… serves as the primary vehicle by which people communicate and record information" (Grishman 1). engineers and scientists aim to adapt computers to our needs, to instill in them the capacity for inference and imprecision which we so readily possess and wield. From this endeavor emerged the field of computational linguistics, "the study of computer systems for understanding and generating natural language" (4). While in a sense all practices of linguistic science seek to deconstruct and fully understand language, most such studies are themselves grounded in human communication and therefore susceptible to its logical shortcomings. Computational linguistics alone[1] looks to break down language not merely to the level of systematic *human* comprehension, but to the level to systematic *mathematical* comprehension, the point at which all uncertainties are resolved into testably exact values. "By understanding language processes in procedural terms, we can give computer systems the ability to generate and interpret natural language" (Grishman 1).

[1] There is another field, known as theoretical linguistics, which is closely related to computational linguistics and so shares many of the traits mentioned here. The theoretical and computational varieties are almost identical fields save for their goal: the former seeks a purely scientific knowledge of language structure, while the latter intends to form practical developments from its research. "Engineering and scientific objectives, of course, usually go hand in hand" (Grishman 6) and so the two could easily be views merely as different aspects of the same field. As such, much of what is discussed in this discourse can be applied equally well to both. However, for simplicity's sake and given the technologically oriented nature of the topic and much of the supporting evidence, I will use the phrase "computational linguistics" with the implicit understanding that I do not mean to exclude theoretical linguistics from consideration.

The obstacle to this goal is that, simply speaking, science is not yet able to analyze human language as a fully procedural process. Before a computer, designed by a human being, can understand natural language, we humans ourselves must understand it – not simply how to make use of it, but how it actually works. In order to do so, computational linguists address the issue of "modularity: dividing our system's knowledge into relatively independent components. Dividing the problem allows us to attack the subproblems independently (or nearly so)" (Grishman 7). Separating natural language processing into a finite number of distinct units with a finite number of known interactions is the first step towards a fully formal understanding of linguistic structure.

Taking this first step, though, is no small task. Most attempts aim to separate the analysis of individual sentences from the structure of the discourse in which they are situated. However, most people do not realize the level to which contextual information affects their understanding of seemingly unambiguous phrases. Consider the simple adage "time flies like an arrow."

> Susumu Kuno of Harvard… asked his computerized parser what the sentence "Time flies like an arrow" means. In what has become a famous response, the computer replied that it was not quite sure. It might mean (1) that time passes as quickly as an arrow passes. Or maybe (2) it is a command telling us to time the flies the same way that an arrow times flies; that is, "Time flies like an arrow would." Or (3) it could be a command telling us to time only those flies that are similar to arrows; that is, "Time flies that are like an arrow." Or perhaps (4) it means that the type of flies known as "time flies" have a fondness for arrows: "Time-flies *like* arrows." (Kurzweil)

Humans hearing this sentence have little trouble understanding its meaning. We know that it makes no sense for an arrow to be timing flies, or to compare flies and arrows, and we have never heard of a species called time-flies, and doubt that such creatures would have a specific preference for arrows. We are instinctively able to synthesize all this information, drawn from outside the sentence, to reach the correct analysis of the phrase itself. But the computer parser, lacking both these data and the knowledge of how to process them, cannot reach that conclusion.

Of course, many of these problems arise from the specific grammatical ambiguities of the English language which crop up in parsing this sentence, and it is true that the same sentiment expressed in other natural languages is not necessarily as difficult for a computer to analyse[2]. However, similar vagueness of meaning is endemic to all natural languages, up to and including those which fell out of common use long before the concept of computers or machine comprehension ever existed. Take, for example, the opening lines of Vergil's *Aeneid*:

   Arma virumque canō, Troiae quī prīmus ab ōrīs
Ītaliam fātō profugus Lāvīnaque vēnit
lītora - multum ille et terrīs iactātus et altō
vī superum, saevae memorem Iūnōnis ob īram,
multa quoque et bellō passus, dum conderet urbem
īnferretque deōs Latiō – genus unde Latīnum
Albānīque patrēs atque altae moenia Rōmae. (Vergil 1-7)

   I sing of arms and a man, who first from the beaches of Troy
went towards Italy and the Lavinian shores exiled by fate –
that man buffeted greatly both on the lands and on the sea by
the force of the gods, because of the unforgetting wrath of savage Juno,
and also enduring many things in war, until he founded the city
and brought the gods to Latium – whence are the Latin race
and the Alban fathers and the high walls of Rome.

This section is an uncontroversial one among classicists, its translation and meaning essentially undisputed among academic circles. However, a computer parser attempting to interpret it would

[2] "El tiempo vuela como una flecha," the same saying translated into Spanish, is essentially unambiguous, largely due to Romance language's far more complex conjugation of verbs compared to that of English. As such, the grammatical ambiguity of "time and "flies" potentially being nouns or verbs does not exist.

run into a number of problems. In the first line, “Troiae” is understood as a possessive genitive attached to “ab ōrīs” – “from the beaches of Troy.” But the form could also syntactically be a dative of direction, a meaning which is less common but certainly not outside of the realm of possibility[3], in which case the phrase would mean “from the beaches to Troy.” Anyone familiar with the tendencies of the Latin language or the story of the *Aeneid* would know that this makes no sense within the logic of the discourse, the whole poem, but viewed out of context, as a simple computer parser would, there is no way to determine the relative validity of one over the other. Similarly, in lines 6 and 7, the subordinate clause begun by “unde” (“whence”) contains no explicitly stated verb. It is common to omit forms of *esse*, “to be,” in cases such as this one, but for a computer to account for this option would mean taking into account the possibility of inserting some form of that verb into every space and testing the result’s grammatical viability. Again, though, a human familiar with this practice of omission or with the story portrayed by the poem would have little difficulty identifying the missing form.

So are humans simply superior to computers in understanding language? Not necessarily. Every language considered so far has been a natural one, meaning it came to be and evolved through human use. Given this origin, it is hardly surprising that humans would be well equipped to utilize it, far better at least than computers are. However, not all languages are natural, and human superiority breaks down when considering artificial languages which are designed for computer comprehension and whose primary objective is not convenience, but precision. A human would need highly specialized knowledge and skill to read basic machine code, consisting of a string of zeroes and ones, and would still be fair inferior to a computer at

[3] In fact, this very syntax is used later in this passage: “Latiō,” meaning “to Latium.” While Latin does contain several syntaxes which are rare to the point of pathology, this usage, and many others, exist in the middle range between common occurence and irrelevance. As such, they would pose a significant problem for computer-based analysis attempts, which have no useful understanding of how to incorporate the rarity of a different syntaxes into translations.

converting that data into information. While higher-level programming languages are designed with the explicit purpose of being easily comprehensible to its human users, the computer which parses a block of code is able to understand it massively more quickly than the person who wrote it.

Given this difference between natural and artificial languages, it is tempting to view the split between the two as a fundamental dichotomy in linguistic classification. However, this distinction is not a result of some inherent difference in the structure of human and computer languages, but rather emerges from the separate needs of these two separate groups, the separate needs which each category came about to fulfill. As such, a language which is meant to be easily understood by both man and machine, would bridge this divide. Take, for example, the Inform 7 programming language, designed to approximate the readability of plaintext English while still being computer-parsable. The statement "a subject is a kind of thing" (Short) is comprehensible both to a human English-speaker and to a computer running an Inform 7 parsing program. Even a more complex declaration, such as "Fitting is a scene. Fitting begins when Wondering ends. When Fitting begins: change conversation set to Table of Fitting Remarks; change the target subject to Theo; now marriage suggests Theo; change the current action to 'Nervously'" (Short), while sounding more unusual to the ear, can still be understood by both person and program.

However, given the current developmental stage of computational linguistics, it is inevitable that the Inform 7 language cannot maintain this level of lucidity through every possible declaration in the language, and indeed it begins to lose its dually-comprehensive coherence in more complex statements. And ultimately, of course, it is a programming language, so when its creator had to choose between clarity to man or to machine, it was inevitably the programmers, and not the program, who lost out:

```
To decide whether (next subject - a subject) produces conversation:
        repeat through conversation set
        begin;
                if the starting entry is current subject or starting entry is blank
                begin;
                        if the final entry is next subject or final entry is blank, yes;
                end if;
        end repeat;
        no. (Short)
```

This statement, while not utterly meaningless to the human mind (and in fact fairly easily recognizable to a programmer as a Boolean function containing a loop), is a far cry from the relative simplicity of natural English. Clearly, this is not the totally comprehensible holy grail of languages which computational linguistics seeks, but it is an early step on the right path. Inform 7 is a language which, though intentionally structured and designed, possesses a level of clarity to human readers which shows it to be spanning the gap between natural and artificial languages.

The existence of 'hybrid' languages[4] such as Inform 7 shows that the divide between human- and machine-oriented languages, between languages of convenience and of precision, is epistemological rather than ontological. As such, the atomization of meaning which is present in and in fact vital to programming languages, which focus on computer comprehension above ease of human understanding, must equally be a component of natural languages. Indeed, the discipline of computational linguistics exists with the ultimate goal of identifying these unambiguous, perfectly precise atoms of meaning, for reducing thus a natural language statement

[4] In the course of this discussion I use the terms 'natural' and 'artificial' to refer to two categories of languages. While these words technically refer solely to the origin of the language – whether it came about through evolution or design – I occasionally use them instead to describe the entity whose comprehension the language most supports: natural for human beings, artificial for machine parsers. Keep in mind that phrases such as "hybrid language" or similar are not intended to suggest that the language both evolved and was designed, but rather that its clarity to people and to computers are closer to equal than is usual among languages, both natural and artificial.

to perfect precision in this way is exactly the step which would allow a computer to parse it and understand the information it conveys.

Furthermore, these linguistic atoms are not some nebulous, theoretical concept, but rather actually an example of a new but well-established scientific theory: memes. Dawkins, the father of memetics, defines a memes as "unit[s] of cultural transmission… tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches" (Dawkins 192). Just as genetics is the basis of biological transmission, so too is memetics the fundamental foundation of informational transmission. As such, memes must, like genes, be precise and unambiguous: though the color of a person's hair or the shape of their face can take on one of a nearly infinite myriad of options, any such quality must necessarily be the result of the presence or absence of a finite number of genes in the individual's DNA; similarly, the apparent fluidity and seemingly unbounded permutations of variety of cultural expression must, in theory, be capable of being broken down into a finite sequence of exact informational atoms. And what better example of cultural information than language, the very method by which we actually transmit such information, the "primary vehicle by which people communicate and record information" (Grishman 1)? Natural languages are inherently no different from any other cultural phenomena: "a vast complex of memes, interconnected and co-evolved" (Blackmore). Language can thus be described as an interdependent collection of informational atoms, known as memes.

This description seems very similar to the way in which computational linguistics approaches the task of breaking down language into its component elements. This similarity is not a coincidence; in fact, it is the very crux of an important realization concerning the nature of the field of computational linguistics: the base particles of meaning, which this scientific discipline seeks to understand, are the memes which constitute the meme-complex of natural language. As such, computational linguistics can actually be understood to be an offshoot of the study of memes, a sort of linguistic memetics. Furthermore, the process of analyzing human language to discover its component memes is essentially a kind of meme-sequencing[5], a quest to unlock the linguistic 'memome'[6], so to speak. Far from being an esoteric field of a neglected science, computational linguistics is on the cutting edge of our developing understanding both of technology and of ourselves.

Language "has the potential for expressing an enormous range of ideas, and for conveying complex thoughts succinctly," and "the aim of computational linguistics is, in a sense, to capture this power" (Grishman 1) through natural language processing. This task is, at its heart, an attempt to analyze and identify the most fundamental base components of language, components which must by their nature be exact and unambiguous in their meaning, and therefore far more easily understood by computers than is natural language in its raw form. Moreover, these basic linguistic elements are actually memes, the informational units of which language is composed, and so therefore natural language processing is essentially meme-sequencing, the act of breaking down an informational meme-complex into its constituent memes. Realizing that information has discreet units allows the science to memetics to be applied as a part of the effort of natural language processing. For example, memes undergo natural selection and evolution just and genes do, and so computational linguists could approach their undertaking by analyzing not only human language as it currently exists, but also as it changed throughout history, perhaps thus more easily identifying the memes which evolved along with it. Even regardless of any potential practical applications, this unification of computational linguistics and memetics is startling in that it provides further more evidence that

[5] As analogous to gene-sequencing.
[6] As analogous to the genome.

languages, and the memes from which they are formed, are living organisms not merely metaphorically but truly, and that our current understanding of the definition of life has only scratched the surface.

Works Cited

Blackmore, Susan. “Memes Shape Brains Shape Memes.” *Behavioral and Brain Sciences* 31.5: 513. Web. 28 Sept. 2009. <http://journals.cambridge.org//?jid=BBS&volumeId=31&issueId=05>.

Dawkins, Richard. *The Selfish Gene*. 30th Anniversary ed. New York: Oxford UP, 2006. Print.

Grishman, Ralph. *Computational Linguistics: An Introduction*. New York: Cambridge UP, 1986. Print.

Kurzweil, Raymond. *The Age of Intelligent Machines*. Cambridge: MIT P, 1990. N. pag. *KurzweilAI.net*. Web. 26 Oct. 2009. <http://www.kurzweilai.net//.html?printable=1>.

Short, Emily. *Glass*. N. pag. *Inform7.com*. Graham Nelson, 2006. Web. 9 Nov. 2009. <http://inform7.com////.html>.

Vergil. *Vergil’s Aeneid: Books I-VI*. Comp. and trans. Clyde Pharr. Rev. ed. 1964. Wauconda: Bolchazy-Carducci, 2007. Print.