

1. What exactly does “word sense disambiguation” mean? How is WSD different from semantic technology?

WSD (or word sense disambiguation) is a critical subset of semantic technology and artificial intelligence. Solving WSD means that computers can understand the meanings of words. That means that computers can understand language. And, language is the key building block of “artificial intelligence” at least if you’re talking about computers that can reason. Accurate WSD has dramatic commercial applications including to significantly improve Internet search and as well as the viability of machine translation and speech recognition.

Specifically, WSD is the task of identifying the correct meaning of a word in context, a task made difficult by the fact that words have so many possible meanings. For example, is “JFK” a president, an airport or a movie? Is “apple” a computer, a fruit, or a record label? Is a “jaguar” an animal, a car, or a football player? When I say “cut”, is that a verb (as in, to sever), or is it a noun (as in, a track of music)? Of course, the reason I was able to pose these questions the way I did is because I provided no context for the words in question. Had I used any of the words in context (for example, “my flight leaves from JFK”), you would have had little difficulty identifying the correct sense of an otherwise ambiguous word. Not so for computers. Until now, computers have been flat-out bamboozled by word senses. That is the problem Idilia has solved.

Let me illustrate in more depth. Take, as an example, the sentence “Tiger Woods left his titanium driver in the clubhouse”. For a human, the meaning of this sentence is pretty easy to decipher. For a computer, doing the same is a gargantuan task. Here’s why:

- Consider the word “Tiger”. This word has at least eight possible meanings, some of which are proper nouns and some of which are common nouns. I assume you’re familiar with the town in Georgia named Tiger? Idilia’s software is.
- Now consider “Woods”. First, a computer has to accept that this could be the plural of “wood” as well as the proper noun “Woods”. In all, there are at least 21 possible meanings for “wood” or “Woods”. By the way, there are an additional four possible meanings for the combination, “Tiger Woods”, including a book written by Andrew Gutelle.
- So, we’re only two words deep into this sentence, and already, there are 172 possible interpretations of “Tiger Woods”.
- The word “left” has noun, verb, adjective, and adverb senses—in fact, 25 senses in all.
- If we complete this analysis of each word in the sentence, the total number of possible interpretations of this sentence a computer must sift through is a staggering 18.5 BILLION. And, that’s just for one sentence, let alone an entire document.

2. You say Idilia’s technology is a scientific breakthrough. Describe the specific breakthroughs Idilia has accomplished.

For the past fifty years, no technology has been able to achieve greater than 60-70% accuracy in word sense disambiguation. This level of accuracy provides almost no additional value beyond selecting the most common word sense in every instance, and therefore does not provide any incremental benefit to existing software capabilities. Idilia’s WSD technology has decisively broken through this ceiling and is 85-90% accurate—within 5-10% of expert human performance. This means that Idilia has closed 80% of the gap between humans and machines for understanding language.

I will describe five of the specific breakthroughs we have accomplished. Please bear in mind that these represent only a selection of the inventions embodied in the 150 person years of R&D that took place over 7 1/2 calendar years to build Idilia’s system.

NUMBER 1—Learning semantic relations

Idilia’s system incorporates multiple linguistic algorithms and linguistic processing components (including many new algorithms invented by Idilia), each representing significant research breakthroughs, emulating the various techniques used by humans to perform WSD. Key among these, Idilia invented an algorithm to analyze and acquire semantic relations (preferences and restrictions), directly from sense-tagged text, which are then used to disambiguate with high accuracy. This algorithm learns, for example that only an animate entity can “eat” in the sense of consume food, but that only an inanimate entity can “eat” in another sense, to corrode. This information represents general knowledge about the world, and is learned automatically by the algorithm as it reads text.

NUMBER 2—Dispute resolution

The overall design of the Idilia system also represents a significant set of inventions. The system architecture supports multiple components, operating in parallel and contributing a homogeneous, consistent set of linguistic features into a common workspace. To enable multiple components operating in parallel, the system includes a second level of automated dispute resolution algorithms (“super algorithms”) to arbitrate between different individual algorithms. For example, when analyzing the sentence “When playing golf, the driver drives the golf cart close to the tee,” an algorithm which analyzes meaning based on topical relationships (i.e. “what is this sentence about”), would suggest the meaning of “drive” most consistent with the topic of the sentence—“golf”—that being, “hit a golf ball” (which would be wrong). However, an algorithm using semantic restrictions would recommend eliminating this sense of “drive” because it is not consistent with the direct object “golf cart”. The super-algorithms mediate between these individual algorithms and select the best answer.

NUMBER 3—Integrated disambiguation of proper nouns

Idilia’s technology represents a breakthrough in disambiguating proper nouns (“named entities”) in an integrated system that doesn’t require any pre-processing, human tagging, or meta-data when disambiguating common words and proper nouns simultaneously in the same document.

Idilia’s embedded named entity recognition (NER) technology is capable of automatically recognizing approximately 500 categories of named entities, or proper nouns. Idilia’s WSD system is the first to seamlessly integrate NER. Named entity recognition is embedded in the system such that proper noun senses are treated as ordinary word senses, allowing WSD to be directly applied to the problem of named entity recognition.

This breakthrough was absolutely critical to making WSD technology usable in modern applications, which contain a high percentage of proper nouns (over half of queries, for example, contain proper nouns). Thus, when disambiguating the word “jaguar”, the system considers all possible senses, including proper noun senses—for example, the car or the operating system, and common noun senses—in this case, the animal.

WSD is much more difficult when proper noun senses are considered, raising the bar significantly for any system that attempts to resolve proper nouns simultaneously with common words. Simply adding the named entity meanings to common word meanings, more than doubles the number of meanings the system needs to consider. A fairly ordinary word like “titanium”, which has only one common meaning, has at least 3 additional proper noun meanings (it’s the name of a book, a musical group and a song). Idilia is the first to even attempt WSD with this massive added challenge. This alone added two years to the time it took to research and develop the system.

NUMBER 4—Acquisition of proper noun meanings

In order to learn the individual meanings and precise semantic relations around a massive number of proper nouns, Idilia researched and developed technological breakthroughs to enable automatic acquisition of proper nouns and pertinent semantic information from structured and unstructured Internet sources. This has significantly contributed to Idilia’s massive knowledge base—the largest precise inventory of word senses and semantic relations ever built. This massive linguistic knowledge base contains approximately 20 million nodes (representing over 5 million proper nouns and 200,000 common word senses), connected by over 50 million edges (covering a range of almost 1,000 types of precisely specified semantic relations). By way of comparison, Wikipedia currently contains approximately 2 million articles covering about half the number of concepts.

NUMBER 5—Disambiguation of queries

Idilia has achieved specific research breakthroughs in the disambiguation of queries (short, loose-syntax or no-syntax, word-strings that otherwise confound and confuse natural language processing systems). Accurate disambiguation of text fragments such as queries imposes significant additional demands on a disambiguation system, due to both the lack of context and the substantial syntactic and semantic differences between queries and ordinary document text. The problem was overcome with separately trained versions of key algorithms, and, in some cases, modifications to algorithms themselves,

and of course vast amounts of query-specific training data to train the algorithms. As a result, Idilia has the world's only system capable of even attempting to reliably disambiguate short text fragments, such as queries.

3. How can Idilia's technology improve Internet search?

Major search engines are not currently utilizing robust WSD. This is easily demonstrated by the results delivered when you type words and phrases that have multiple meanings. Consequently, Idilia's technology is meaningfully additive to current search technology. Simply put, it will make search results more accurate, and it will therefore serve as an important competitive advantage for those that implement it.

In general terms, Idilia's WSD technology can meaningfully improve internet search results in three ways: A) it will improve the relevancy of query-based search results, B) it will increase the number of relevant paid-listing ads placed per-page in search results, and C) it will improve the quality and coverage of matching advertisements to web pages.

More specifically, Idilia improves search technology in the following ways:

- Current search engine indexes contain keywords, which are essentially character strings. Idilia's technology allows words to be indexed by their specific sense or meaning as derived from the word's context within the page being indexed.
- Idilia's technology disambiguates queries—that is, selects the correct senses of words in queries (including proper nouns), allowing more precise matching—fewer incorrect matches—when matched to a keyword sense index.
- Idilia's technology paraphrases queries into new queries that may or may not contain any of the original search terms but are nonetheless equivalent in meaning (and in some cases, superior in form of expression), generating additional search results that are precisely on topic, but may not contain the original query terms.
- Idilia's technology identifies precise semantic relations between words in text allowing this additional information to be indexed along with word senses and matched to equivalent relations captured in queries.

4. By how much can Idilia's technology improve search results? Is that really a remarkable improvement? How do you measure this?

Idilia's WSD enables search engines to match the senses of words in a query precisely to the senses of words in a possible search result. This functionality offers a fundamental improvement in relevancy over current search techniques. In tests utilizing a leading commercial search engine, Idilia's technology improved relevancy by more than 150% for queries for which the initial search engine results were very poor, and by 30% for queries for which the search engine results were of moderate relevancy. (In this case, the measurement was done by evaluating whether the result returned contained the same word senses as those used in the query, or a precisely equivalent paraphrase thereof, in the same relation to one another as expressed in the query.)

FAQ

5. How will this technology improve paid search?

Idilia's technology can materially improve the placement accuracy of both paid-listings and contextual ads by 20-25%. First let's look at paid-listings. In the same way that WSD adds new relevant results to algorithmic search results by paraphrasing, more ads can be matched to more queries - in this case by paraphrasing the query to generate more ad matches. This generates additional relevant ad matches from the search engine's current ad inventory automatically and with no change in behaviour. Optionally, advertisers can bid on word senses, not just words. Sun Microsystems and Starbucks could both bid on the word "java" since they're not actually bidding on the same word any more. Furthermore, the search engine can place ads against more relevant search results. So, ads for "tasty java" won't show up alongside results for software consulting firms. In just the same way as WSD makes algorithmic search results more relevant, (i.e., by matching word senses in queries to correct word senses in documents), matching ads can be made more relevant—in this case by matching the senses of ad keywords to the senses of words in queries.

In contextual ad placement, advertisers can specify the context of web pages they want their ads placed on. In the process of indexing publishers' web pages, Idilia's software can be used to better evaluate the meaning of the content on the page. In an article about "suits" Idilia's technology will ascertain whether the article is about fashion or litigation, and match ads accordingly. In addition, with Idilia's technology, undesirable content can be much more accurately filtered out. WSD identifies where common words are being used in a vulgar or profane fashion—these are simply other word senses the software can choose from. Furthermore, for any given sense, Idilia's knowledge base can provide additional context, by telling you, for example, that word sense is highly pejorative or connotes something criminal.

6. How is Idilia different from Google or Yahoo? How is it different from Powerset or Hakia?

Idilia is a company that develops artificial intelligence technology. Idilia is not a search engine. Specifically, Idilia has developed artificial intelligence technology to enable computers to understand language, specifically to solve the AI problem of automated WSD. WSD is the critical prerequisite for computers to understand and work with natural language—language expressed by humans. Accurate WSD technology will power the search engines of the future. It will also be the technology that ultimately supports the widespread deployment of continuous speech recognition systems—speech recognition systems that perform like people. Accurate WSD will open the multi-billion dollar translation market to viable software applications, because it enables machine translation software to attain commercially acceptable levels of accuracy.

7. How will Idilia's WSD revolutionize machine translation and speech recognition?

Machine Translation is another immediate use for Idilia's WSD technology. The primary reason existing software cannot accurately translate is due to the ambiguity of words, not just in English, but in all languages. If you get the grammar wrong in a translation,

people will forgive you, but if you mix up word meanings—and computers often do, because they don't know which meaning of a word they are supposed to be translating from or to—your translation is unintelligible. For example, translate “the thief left empty handed” to French using Babelfish and translate it back to English and you will get “the left robber empty given”. Idilia's WSD solves the problem of resolving word meaning.

In speech recognition, we should first be clear that today's technology is very good compared to just a few years ago. However, the bulk of that improvement has come in what I'll call the acoustic layer of the software. That is, the software has become very good at recognizing sounds and coping with pronunciation, accents, cadence and the like. But, words that sound the same can have different meanings, hence the need for WSD. The problem is even bigger when you consider that sometimes, entire groups of words sound the same as one another.

If you correctly recognize what someone has said (i.e. you make an accurate transcription), you still need to interpret what it means. This requires attaching word meanings—when I say “cut the lights”, I don't mean “cut” in the same sense as “cut the cake”. I see this application as fundamentally changing how call centers work, for example. Voice response software could actually become as, or even more, efficient than speaking with a human. And, for certain, you'll be able to talk to your car and your phone or PDA. That's a given. In fact, WSD will change the interface between humans and computers, allowing keyboards to be replaced with a voice interface where and when it makes more sense to speak to a computer.

8. Why did it take so long to develop accurate word sense disambiguation, and why hasn't this been achieved before?

A great deal of research has been conducted over the last 50 years to attempt to solve WSD, but none of the prior research has achieved commercially viable levels of accuracy. One reason for the failure to solve the problem until now is that, historically, most WSD research has taken place as single-discipline endeavors or loose collaborations at a distance, so that computational linguists and machine learning experts were not able to leverage each other's skills fully. As it turns out, a tightly integrated, cross-discipline, collaborative approach is needed to solve the problem.

Also, advances in hardware (a frequently pointed-to explanation for many recent technology breakthroughs) have been a factor in our WSD breakthrough. Since there are no shortcuts in WSD R&D, advances in hardware have recently shortened the cycle time of experiments and therefore led to more rapid gains in performance. If Idilia were using the same hardware, particularly CPU and memory configurations we were using three years ago, I would hazard to guess, armed with a known R&D methodological approach, even Idilia would still be two years away from developing the commercially deployable WSD technology we have today.

We are extremely proud of the range of research breakthroughs we achieved on a variety of fronts, and many of our individual algorithms embody major inventions. We invented algorithms for extracting and using common sense semantic relations from text, for example, that represent, in and of themselves, a major research milestone.

9. What patents support Idilia's technology?

Idilia has filed five patents and continues to file additional patents and continuations on existing patents. For the time being, we have a major patent pending for key inventions incorporated in our Word Sense Disambiguation system. We also have more specific patents pending concerning Internet search. We have a patent pending concerning specific inventions regarding the disambiguation of queries, including disambiguation of queries on mobile devices, and we also have patents pending against key aspects of the application of our word sense disambiguation and paraphrasing technology to Internet search and advertising. Lastly, Idilia has recently filed a patent for its research breakthroughs in knowledge extraction—the automatic mining and linking of new word senses and terminology into a general-purpose semantic knowledge base.

10. How much time has Idilia committed to R&D to build its technology?

Researching and developing accurate WSD is a long and intensive process involving a great deal of pure research. There are really no shortcuts. We have spent more than 7 1/2 years (120 person years) developing deployable, commercially robust, WSD technology. During this time, Idilia has remained solely focused on the R&D of WSD, and has deployed highly skilled resources as rapidly as possible given the complex and integrated nature of the problem. This technology now encompasses more than 1.5 million lines of C++ code, and numerous significant and patentable inventions. We believe that even a much larger company would require a similar amount of time to conduct this research and development.

11. Who are Idilia's investors?

Idilia has been backed by The Hearst Corporation, Canada's National Research Council, as well as private investors and venture capital investors.

12. Can Idilia's technology be readily deployed for large scale, real time applications?

Yes, Idilia has developed products incorporating its WSD technology, with comprehensive APIs, that can be deployed in large-scale applications across thousands of CPUs. The technology has been stress-tested. Idilia's technology will be capable of disambiguating a typical search query in less than 1/10 of a second in large-scale deployment, and all Idilia's products can be deployed on conventional 64-bit hardware.