Micro Reading with Priors: Towards Second Generation Machine Readers

Ndapandula Nakashole
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
ndapa@cs.cmu.edu

Tom M. Mitchell
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
tom.mitchell@cs.cmu.edu

.

Abstract

Micro reading in machine reading can be compared to deep reading in human reading. Deep reading has been defined as a set of processes that enable comprehension and that include inferential and deductive reasoning, analogical skills, critical analysis, reflection, and insight. In this paper, we sketch what we envision to be a viable approach to micro reading. The proposed approach leverages the *knowledge* that has been acquired by the machine readers that have been developed to date.

1 Introduction

Machine reading systems have so far achieved a degree of success through *macro-reading* of relational facts. Macro-reading is a *shallow* way of machine reading which leverages the redundancy of huge corpora to capture language patterns. Such patterns are then used to identify facts expressed text. Macro-readers lack the ability to do *micro-reading* — the full comprehension of a single instance of discourse, for example, a document, paragraph, or sentence and being able to answer comprehension questions about exactly what is expressed in that single document, paragraph, or sentence.

To micro read non-trivial pieces of text, adult readers engage in what is referred to as *deep reading*. Central to deep reading is inference, which involves drawing upon *prior knowledge* about the concepts involved. Studies of brain scans of people's brains while reading fiction have found that readers mentally simulate each new situation encountered in a story[4, 24]. Details about actions and sensation are captured from the text and integrated with personal knowledge from past experiences.

While brains of adults seamlessly engage in deep reading, micro-reading in machine reading is still in its infancy and has been relatively under-explored in comparison to macro reading. In this paper, we propose an architecture that tightly integrates micro reading into machine reading.

Our proposed approach is not to replicate the little understood processes of the human brain. In recent years, much effort has gone into developing methods for knowledge acquisition [1, 3, 21, 23]. The resulting methods have produced a wealth of knowledge. This knowledge ranges from clean but limited coverage knowledge bases, to huge corpora of text-level assertions with high coverage but also a non-negligible amount of noise. Other kinds of knowledge is available through lexical resources and linguistic treebanks. All of this knowledge can now be brought to bear in an effort to develop machine readers capable of reading at a more advanced level (micro readers) than the first generation (macro readers) have been able to do. This mimics, albeit in a primitive manner, how humans become increasingly capable of deep reading due to knowledge and experience acquired over time.

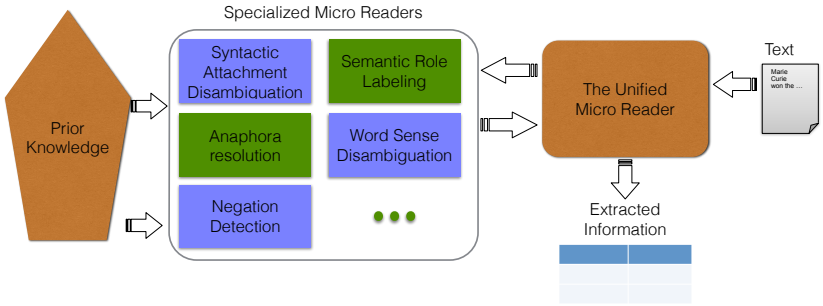

Figure 1: A sketch of the proposed approach of micro readers whose annotations are utilized by the unified micro reader.

The need for prior knowledge in advanced machine reading has long been acknowledged. For example, the ongoing DARPA program named Deep Exploration and Filtering of Text (DEFT) [6] is facilitating research working towards realizing this goal. In this paper we sketch what we see as a viable approach towards prior-knowledge aware micro readers.

We propose an architecture of specialized micro readers, each addressing a clearly defined aspect of language understanding. Each specialized micro reader produces annotations of the input text. A single unified micro reader uses these annotations to produce the most likely interpretation of the input text. To illustrate the concept of specialized micro readers, we will later explore one such micro reader, which addresses a special type of syntactic ambiguity [1].

We could, for example, build micro readers for: syntactic attachment disambiguation (for instance due to prepositional phrase attachment ambiguity), word sense disambiguation, negation detection, anaphora resolution, semantic role labeling, and others [2].

Challenges. Our proposed modular approach is motivated by the challenges that micro reading entails:

- **Relevant Knowledge Identification:** Which knowledge facilitates better language understanding? We conjecture that it is easier to answer this question when the natural language aspect we want to understand is highly specific. For example, the knowledge required for solving syntactic ambiguity is different from the knowledge required for negation detection.
- **Utilizing Knowledge:** Having identified the relevant knowledge, how do we utilize this knowledge for the task at hand? In other words, what models are suitable for incorporating prior knowledge. Here again, which model is better is likely to be a task-specific answer.
- **Diversity of Knowledge:** The type of knowledge that facilitates micro reading is diverse and goes far beyond relational knowledge. There is relational knowledge, there is knowledge about language rules, their is lexical knowledge etc. Due to this diversity of knowledge, how does one weigh potentially conflicting knowledge? How do we decide which knowledge is more informative. Here again, which type knowledge is more informative is likely to be a task-specific answer.

The above challenges motivate our proposed divide-and-conquer approach.

2 Micro Reading Architecture

A sketch of the proposed architecture of specialized micro readers is depicted in Figure 1. As input, the system takes text, such as a sentence or an entire document. The unified micro reader passes

[1] When a reader can reasonably interpret the same sentence as having more than one possible structure, due to many possible relationships between pairs of words, the sentence exhibits syntactic ambiguity.

[2] An interesting questions is how many such micro readers are sufficient for reading. The answer probably depends on the desired level of understanding.

the text on to the specialized micro readers, which return annotations of the text in one common annotation language. These annotations enable the unified micro reader to read each sentence better than it would have been in the absence of prior-knowledge aware micro readers. Notice that the unified micro reader itself can make use of prior knowledge as it decides the final interpretation of the text. In the figure we illustrate the concept of specialized micro readers with a few such micro readers. Any number of micro readers can plugged into this modular architecture. The integrated unified micro reader is then faced with the non-trivial task of generating the most likely interpretation of the text, given the different annotations. Notice that here the unified micro-reader is, in principal, doing joint inference.

To see how knowledge might be helpful, let us consider the following examples:
Co-reference resolution: "The bee landed on the flower because it wanted pollen."' If we know that the bees feed on pollen, we can correctly determine that "it" here refers to the bee and not the flower.
Negation detection: "Things would be different if Microsoft was headquartered in Texas." From this sentence alone, a fact extractor might extract that Microsoft is headquartered in Texas. But from the prior knowledge that Microsoft was never headquartered in Texas, we might be able to better detect the negation here, in addition to the syntactic cues such as "if".

We now analyze one micro reader in detail, the one for resolving a specific type of syntactic attachment ambiguity.

3 Syntactic Attachment Ambiguity

Syntactic ambiguity occurs when one sentence can be interpreted in more than one way due to ambiguous sentence structure. It occurs not from the range of meanings of single words, but from the relationship between the words and clauses of a sentence, and the sentence structure implied. When a reader can reasonably interpret the same sentence as having more than one possible structure, e.g., possible different relationships between words, the sentence exhibits syntactic ambiguity.

A common cause of syntactic attachment ambiguity is *prepositional phrase attachment ambiguity, (PPAA)*. For example, consider the following sentences:

1a. Alice caught the butterfly with the spots

1b. Alice caught the butterfly with the net

2a. The government discovered irregularities in the adoption process.

2b. The government discovered irregularities in June.

In sentences 1a and 2a, the prepositional phrases (*with the spots*, and *in the adoption process*) attach to the *nouns*. If the task at hand is relation extraction, we get *binary* extractions of the form:

4a. ⟨Alice⟩ caught ⟨the butterfly with the spots⟩

5a. ⟨The government⟩ discovered ⟨irregularities in the adoption process⟩

However, in 1b and 2b, the prepositional phrases (*with the net*, and *in June*) attach to the *verbs* in these sentences. If the task at hand is relation extraction, we get *tenary* extractions of the form:

4b. ⟨Alice⟩ caught ⟨the butterfly⟩ with ⟨the net⟩

5b. ⟨The government⟩ discovered ⟨irregularities⟩ in ⟨June⟩

Therefore, prepositional phrase attachment disambiguation addresses the question of: Given a sentence where there is a sequence of words with the following part-of-speech tags: *V, N1, P, N2* , does the prepositional phrase (P N2) attach to the noun (N1) or the verb (V)? Prior work has addressed this problem with statistical approaches, primarily based on co-occurrence frequencies of words. [5, 9, 16, 17, 20, 22]. The problem is cast as binary classification to determine if the attachment site of the prepositional phrase is the noun or verb. Other works cast this problem as dependency parser correction problem. That is, given a dependency parse of a sentence, with potentially incorrect prepositional phrase attachments, change it such that the prepositional phrases attach to the

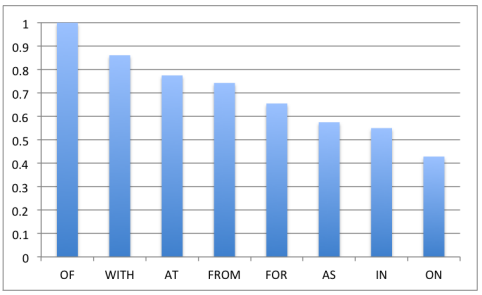

Figure 2: Accuracy of Stanford parser for the top eight most frequent preposition on PPAA instances from news data.

correct words. In all these works, no prior knowledge is taken into account, except for semantic type information [20].

Most machine readers handle PPAA through a dependency parser. In other words, the machine readers leave PPAA to the dependency parser. We therefore carried out an experiment to access performance of a state-of-the-art dependency parser (Stanford parser) on the PPAA problem. We automatically identified a number of sentences containing PPAA sequences, *V, N1, P, N2*, from a recent news corpus and applied the Stanford parser to these instances. We manually evaluated 40 attachment instances per preposition for the eight most frequent prepositions in the corpus.

The outcome of this experiment is shown in Figure 2. While there are some prepositions where the parser does well, for example "of" (which tends to occur with noun attaching PPs), for some prepositions, the parser's precision is around 50%. Thus there is still room for improvement when it comes to PPAA and this is where the potential of prior knowledge can play a role. Drawing a distinction in performance of different prepositions is on notable importance. We say this because, the problem of PPAA may be perceived as a solved problem due to the established baseline [5] scoring quite high on the Wall Street Journal, corpus, 84 % precision. However we re-implemented and reproduced the baseline solution, and found that, if we remove the preposition "of" [3], precision drops to only 78%.

Since prior approaches rely on co-occurrences of sequences of words, it is clear that sparsity can become an issue. That is, if we encounter a sentence whose sequence of words was not part of the training data, the best one can do is to approximate from less informative co-occurrences of sub-sequences such as in the back-off model [5]. In contrast, leveraging prior knowledge about individual words and their corresponding real world concepts can help alleviate sparsity issues.

Suppose we have prior knowledge in the form of a simple proposition that says a butterfly can have spots, in the form of a subject-verb-object: *butterfly-has-spots*. We can use this information to infer that a noun attachment is possible in *1a*. Similarly suppose we have another proposition *process-marred_by-irregularities*. Knowing that a process can be inflicted with irregularities can help us decide that a noun attachment is possible in *1b*. Thus we can see that each preposition can be mapped to relevant knowledge base relations, (verb phrases in this case) , that indicate an attachment one way or the other. The details of how this mapping can be achieved is beyond the scope of this paper, however, our ongoing work has shown the feasibility of this mapping. Relational knowledge is just one type of knowledge, other types of knowledge can be leveraged in this manner.

4 Discussion

4.1 Logical Forms and other Semantic Representations

Semantic parsing is aims to explicitly annotate text with its semantics. In the same way syntactic parsers annotate text with its syntax such as part-of-speech tags, and dependency relations. There-

[3] In this established baseline, and in the majority of PPAA solutions, "of" is by default attached to the noun.

fore, a central goal of semantic parsing is to transform natural language into representations that that can be easily executed by a computer program to, for example, answer questions.

Montague semantics, in the form of Lambda Calculus, has become a common representation in semantic parsing. On example of using Montague semantics is Combinatorial Categorical Grammar (CCG) [19], which has been adopted by many semantic parsers [26]. Liang et al. [10] introduced another kind of semantic representation for compositional semantics, known as dependency-based compositional semantics (DCS). In DCS, the logical forms are trees, in the manner of syntactic dependency trees in the realm of dependency parsing.

Banarescu et al. [2] have introduced a semantic representation language called Abstract Meaning Representation (AMR). The AMR project aims to produce a sizable sembank with thousands of sentences, manually annotated with their semantic meanings. The authors believe such a resource will facilitate the development of widely usable semantic parsers, in the same way the Penn Treebank has facilitated the development of widely used syntactic parsers. AMRs are rooted, directed, edge-labeled, leaf-labeled graphs.

Our ongoing work involves determining the semantic representation best suited to our vision of micro-reading, which could be one of the above or a different one.

4.2 Low Dimensional Vector Space Embedding

Low dimensional vector space representations of language constructs such as words and phrases is increasingly becoming a common way to overcome sparsity [11, 12]. We have began clustering verb phrases that appear in SVOs. We have also begun clustering entities. We currently use Principal Components Analysis (PCA) for dimensionality reduction but we can also leverage new developments in dimensionality reduction.

4.3 Connection to Knowledge on-Demand (KoD)

Recall that the first step in developing a specialized micro reader is to identify the relevant know. Let us assume we can perform this task reasonably well. This means that for a given instance of a micro reader, we know precisely which question we want answered before we can make a decision. This means that when a static knowledge base suffers a recall problem and does not contain the information we want, we can leverage KoD services by sending targeted queries, such as the OpenEval [18] KoD service in NELL.

5 Conclusion

The premise of this vision paper is that, we have only scratched the surface with prior knowledge as a vehicle for advanced machine reading. With the amount of knowledge that has been accumulated by first generation machine readers and information extractors, we are now in a position to develop the second generation of machine readers. We proposed what could be a viable high-level architecture for prior-knowledge aware micro reading.

Acknowledgments

This work was supported in part by DARPA (award number FA87501320005), and Google. Any opinions, findings, conclusions and recommendations expressed in this papers are the authors' and do not necessarily reflect those of the sponsors.

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z.G. Ives: DBpedia: A Nucleus for a Web of Open Data, ISWC/ASWC 2007

[2] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider: Abstract Meaning Representation for Sembanking. In Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 178186, Sofia, Bulgaria. 2013.

[3] A. Carlson, J. Betteridge, R.C. Wang, E.R. Hruschka, T.M. Mitchell: Coupled Semi-supervised Learning for Information Extraction, WSDM, pp. 101-110, 2010

[4] N. Carr:The Shallows: What the Internet Is Doing to Our Brains. W.W. Norton, 2010

[5] M. Collins, J. Brooks: Prepositional Attachment through a Backed-off Model. Third Workshop on Very Large Corpora, pp. 27-38, 1995.

[6] DARPA: Deep Exploration and Filtering of Text. `http://www.darpa.mil/Our_Work/I2O/Programs/Deep_Exploration_and_Filtering_of_Text_(DEFT).aspx`. Retrieved,:Oct, 2014.

[7] A. Fader, S. Soderland, O. Etzioni: Identifying Relations for Open Information Extraction, EMNLP, pp. 1535 - 1545, 2011

[8] L. Fang, A. Das Sarma, C. Yu, P. Bohannon: REX: Explaining Relationships between Entity Pairs. PVLDB 5(3), pp. 241-252, 2011

[9] D. Hindle, M. Rooth: Structural Ambiguity and Lexical Relations [find similar] [try Google] Computational Linguistics, 19(1), pp. 103-120, 1993.

[10] P. Liang, M. I. Jordan, D. Klein: Learning Dependency-Based Compositional Semantics. ACL 2011.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean: Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

[13] T. Mohamed, E.R. Hruschka, T.M. Mitchell: Discovering Relations between Noun Categories, EMNLP, pp. 1447-1455, 2011

[14] N. Nakashole, T. Tylenda, G. Weikum: Fine-grained Semantic Typing of Emerging Entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1488-1497, 2013

[15] N.Nakashole, T. M. Mitchell: Language-Aware Truth Assessment of Fact Candidates In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1009-1019, 2014

[16] P. Pantel, D. Lin: An Unsupervised Approach to Prepositional Phrase Attachment using Contextually Similar Words. ACL, 2000.

[17] A. Ratnaparkhi: Statistical Models for Unsupervised Prepositional Phrase Attachment ACL/COLING, pp. 1079-1085, 1998

[18] M. Samadi, M. M. Veloso, M. Blum: OpenEval: Web Information Query Evaluation. AAAI 2013

[19] M. Steedman. 2000. The Syntactic Process. MIT Press.

[20] J. Stetina, M. Nagao, J. Zhou, K. W. Church: Corpus Based PP Attachment Ambiguity Resolution 1 with a Semantic Dictionary Fifth Workshop on Very Large Corpora, 1997

[21] F.M. Suchanek, G. Kasneci, G. Weikum: Yago: a Core of Semantic Knowledge, WWW, pp. 697-706, 2007

[22] K. Toutanova, C. Manning, A. Ng: Learning Random Walk Models for Inducing Word Dependency Distributions. ICML 2004

[23] G. Weikum, M. Theobald: From information to knowledge: harvesting entities and relationships from web sources. PODS 2010

[24] L. Wehbe, A. Vaswani, K.Knight, T. Mitchell :Aligning context-based statistical models of language with brain activity during reading. EMNLP 2014

[25] M. Wolf , M. Barzillai: The Importance of Deep Reading. Challenging the Whole Child: Reflections on Best Practices in Learning, Teaching, and Leadership, ed. by Marge Scherer. ASCD, 2009

[26] L. S. Zettlemoyer and M. Collins: Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. UAI 2005.