A Machine Learning Approach to Coreference Resolution and Wh-Sentence Tracing.

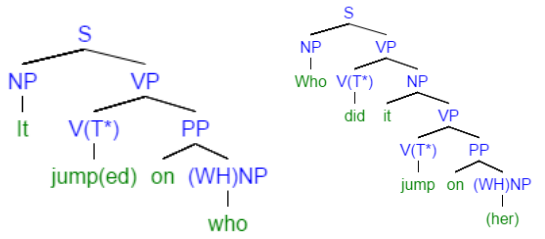
Mariah D. Davis
Michigan State University
229B Emmons Hall, East Lansing MI
davism90@msu.edu

Abstract.

This paper presents a learning approach to coreference resolution of noun phrases in temporal-semantically dependent, wh-sentences. The approach uses a small annotated corpus for training, and determines coreferencing of both nouns to pronouns and nouns to nouns, while not accounting for entity types of noun phrases. In addition, the approach follows the Discourse Prominence Theory, incorporating not only semantic weight towards accuracy, but also syntactic significance. The approach as a stand-alone approach does not seem to present any notably helpful results in coreference resolution; however, when paired with other coreference resolution techniques, it's likely that the approach will help to further improve accuracy in both ambiguous and temporally dependent wh-sentences.

I. Introduction.

Coreference resolution is the linguistic process of tracing one entity to another within a given context. There are many factors which affect coreference in human sentence processing, such as the type of anaphoric expression and the interaction with focus antecedents. In the case of categorical and pronominal coreferencing, read times are significantly faster than in categorical and repeated coreferencing. Faster readings are also obtained from sentences preceded by several sentences in which the coreferenced noun has been explicitly defined. Likely, these faster reading times are a result of lexical processing and alterations of the syntactic structure. Another factor is the saliency and plausibility of the interpretation, which affects coreference resolution within discourse processing. This is particularly noticeable in ambiguous anaphors, where the discourse processing is significantly impacted by the type of quantification of head nouns. Furthermore, it has been theorized that the discourse processing of coreferencing and ambiguity resolution receives an impact from the existence of the coreferenced noun within context. While a coreference pair may exist syntactically, if the entity itself does not exist in a given context, the coreferenced entity by default also does not exist. This semantically false existence seems to override the syntactically correct structure of the

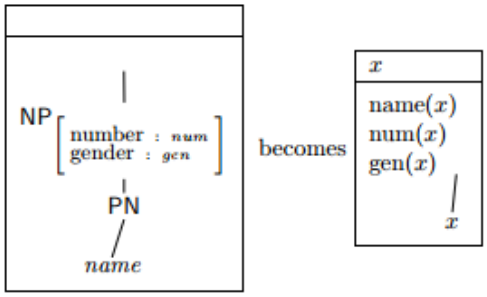


pair and leads to an implicit, if not somewhat abstract falsification of the semantic connection between the two entities.

Wh-movement in sentences is the linguistic phenomenon in which an entity in the canonical position is referenced in the beginning of the sentence, rather than reference overriding the entity's canonical placement. The shifts to the head of the sentence, thus making the canonically positioned entity a head, as well. Such in the sentence "It jumped on her." can be rewritten as "It jumped on who?" to create a question. Yet, more often than not, it is met with the reply "Who did it jump on?", an defined example of wh-movement. Linguists have yet to decide on the cause of the movement, whether syntactic and semantically originated, or solely a semantic process, has yet to be decided.

In wh-sentences, coreference resolution is imposed upon by several factors, including, but not limited to, those previously listed. Additional factors consist of distance, gender matching, a lack of entity to be linked to, and multiple entities

linked to one entity. Returning to the aforementioned issue, wh-movement is also faced with problem of representation. While some linguists believe it to be a syntactical process, in which the wh-word is copied, moved, and referenced to the entity in the canonical position, others believe it to be completely feature based. The wh is introduced as a completely new entity, in this case, and determined by the



addition of features as the sentence progresses. Something similar can be viewed in the diagram taken from (Wiemer Hastings & Iacucci, 2001) in regards to the DRS (Discourse Representation Structure):

This approach to coreference resolution seeks to resolve the factors of gender disambiguation, a missing entity, and multiple entity disambiguation within wh-sentences in a similar manner. This approach additionally follows the assumptions of Discourse Prominence Theory when determining the likelihood of coreference pairs, which states that pronouns are more likely to be coreferenced to entities that have already been established within the context, and that syntactic restraints heavily affect which entities are prominent. Ultimately, the approach seeks out an additional technique to help improve coreference resolution, not to serve singularly as an attempt at it.

II. Related Works.

Somewhat expectedly, the best performing models of coreferencing are those with a heuristic approach (Hobbs, 1978; Lappin & Leass, 1994; Mitkov, 1998), or rather one that is focused less in a psychological approach. Grosz and Sidner (Grosz, 1977; Sidner, 1979) created an approach concentrated on focusing effects (Anderson, Garrod, & Sanford, 1977; Sanford & Garrod, 1981), as mentioned earlier to play a large part in coreference resolution and read times in humans. Centering theory, that which links focus, referring expression, and coherence of utterance within a segment of discourse processing (Grosz, Joshi and Weinstein, 1995), has also been suggested as an approach to establish salience. By combining several aspects of human processing, this approach does well to show that one focus approaches do not solve the issue of coreference resolution. Instead, this is a multilayered issue requiring several aspects to be addressed. (Poesio, 2001) created an approach, followed by later studies in (Poesio & Stevenson, 2002), that evaluate current approaches compared to psychological approaches to coreference resolution.

In (Seidenberg et al., 1984; Tanenhaus & Lucas, 1987; Morris, 1994; Hess, Foss, & Carroll, 1995; Garrod & Sanford, 1994; Gordon & Hendrick, 1998), the separation of discourse and lexical processing, as this paper will attempt to integrate, is discussed. Contrastingly, (MacDonald et al., 1994; Elman, 2009; Levy, 2008) have likened discourse and lexical processing as a uniform approach, and while not accounted for in the approach of this paper, a divide between the uniformity, or lack thereof, should be noted. The processing of separating lexical and discourse while maintaining both structures has led to lexical consistency accompanied by inconsistency in discourse (Rayner et al., 2004; Warren & McConnell, 2007; Warren et al., 2008; Joseph et al., 2008; Hess et al., 1995; Duffy, Henderson, & Morris, 1989; Morris, 1994). Due to the inability discriminate the separation point between lexical and discourse processing, it is difficult to say what weight the approach of this paper will need to give to each lexical and semantic factor addressed to improve coreference resolution. However, in support of discourse, (Moxey & Sanford, 1993, 2000) have addressed quantifier types, in which the lexical structure remains the same, to examine the semantic effect. (Huang & Gordon)'s experimentation further adds to the evidence of discourse and lexical processing separation, in which there was an equivalent delay of processing up to the introduction of context, thus showing, lexically, referring expression are

treated the same. Therefore, an established separation of lexical and discourse processing is the chosen theory for this paper's approach.

Warren and colleagues (2008) addressed the issue of existence of referring expressions and noun phrases within a given context. By alternatively placing plausible vs implausible situations in real and fantasy worlds, it was discovered that read times were much slower for implausible situations in the real world, while in a fantasy world, read times for plausible and implausible situations were similar. This may provide insight as to the weight that existing context and repetition places on coreferencing resolution.

Lastly, the Discourse Prominence Theory, created by (Gordon & Hendrick, 1998) was developed based on a theory called Discourse Representation Theory (Kamp & Reyle, 1993; van Eijck & Kamp, 1997). The DRT accounts for accessibility, and disallows for linking of referring expressions if they are separated syntactically from the main entity. Moreover, the DRT traces features from the main entity to the referring expression and accounts for satisfaction of these features, both processes of which can be found within the approach of this paper.

III. Machine Learning Approach to Coreferencing.

As stated previously, this approach should not be taken as a standalone approach to machine learning. The approach will focus on two main types of sentences; sentences in which the referring expression must be matched to a specific entity in a linguistic unit of multiple entities of differing gender, and sentences in which the referring expression has no stated gender, and must be linked to multiple entities. The approach, as a result, is a dual layered process, where the first layer accounts for matching gender of referring expressions to the main entities, and the second layer accounts for the likelihood of a specific main entity, and its paired, gender-matched referring expression drawn from the collected coreferenced expressions of the first layer, belonging to indexed entity of the sentence. The layers and their calculations are represented in matrices, produced by "Armadillo", a C++ library allowing for the implementation of MATLAB applications. The entire program was written in C++ using 856 lines of code, and Armadillo being the only externally imported library.

The approach is required to use a manually annotated corpus for both testing and training purposes. Due to the purpose of the approach solely being to increase the accuracy of coreference resolution, a POS tagging mechanism was not focused on. However, if need be, one can be incorporated. The current, viable annotations consist of PERSON, PER_DESC, VERB, and ADVERB, where person is a proper or general noun phrase, per_desc is any referring expression, verb consists of verbs, and adverbs refer to temporal expressions, such as "when", "before", "now", and others. The tags are able to be changed, dependent upon the annotator's choice, and are suggested to do so, as they were poorly chosen for this approach, as, nevertheless, the tag names are not particularly crucial to the approach.

The program begins by reading through a document, where the discourse separation occurs in a line break format per first layer analysis, and ceases to exist on second layer processing. Moving into first layer processing, main entities and referring expression are separated into containers by chronological occurrence. The distance is then calculated between the two containers for the occurrence of main entity, following the occurrence of the referring. For each gender of referring expression occurring within the same discourse, a feature weight is added using linear regression in the form of:

$$\hat{y} = \beta_0 + x_1 \beta_1 + x_2 \beta_{2\ldots} + x_n \beta_n$$

Where β_1 is m, in the linear expression y=mx+b, and b= β_0, and so forth to account for the three dimensional linear expression, and x_n represents the features being account for. In this approach, neither β_1, β_2, nor β_0 have a negatively weighted effect, as introduced in the related works, there is an uncertainty about the weight of lexical representation over discourse processing and vice versa. There are no bounds on the linear equation established, as the evaluation of the features will simply be comparative in nature. Furthermore, it was not explored whether the linear regression approach was appropriate for the task at hand, as

it was the first and only method applied, due to time restrictions. It is likely that other methods may demonstrate better results.

In the instance of four sentences:

1. "Mary was good at her job."
2. "Bob liked Mary. She liked him, too."
3. "Bob knew his limitations."
4. "When she was little, the building burned down. It was a basketball court. Now it is a skating rink."

The resulting matrix consisted of:

Referring Expression x Main Entity	Mary	Bob	Building	Basketball court	Skating rink
her	2	0	0	0	0
she	1	1	1	1	1
him	1	1	0	0	0
his	0	2	0	0	0
it	0	0	1	1	1

The program then runs through an algorithm that called for the highest likelihood of main entity-referring expression pairs to be passed into the next layer. This includes "her-Mary", "Bob-his", "it-building", "it-basketball court", and "it-skating rink". The program completely disregards she and him because the gender within the second sentence, the two noun phrases and referring expressions are unable to be determined.

Once passed into the second layer, each pair is then applied to each sentence, regardless of line breaks. The program checks for four factors: temporal initialization, declarative existence triggered by a to-be verb, if the main entity and referring expression are in the sentence, and if they are, the distance between the coreferring expression and the entity, including entity changes preceding or following the referring expression.

All temporal evaluation is trigged by the existence of temporal schema. These are explicitly defined, by words such as the temporal "when". When such a word is found, the program will apply a positive value to the noun-pronoun pairing in the sentence that follows the temporal trigger. The reasoning being, should a temporal exist, with identical referring expressions, it is likely that the temporal is expression the change in the referring expression over two time periods.

Another positive value is applied if a declarative existence is known. The program will evaluate the sentence, and seek a to-be verb. Should one be found, it will match the entity or referring expression before and after the to-be verb, and increase the value of that pairing to the sentence. While this has a chance of being inaccurate, it is likely that the existence of the to-be verb creates a direct link between the noun and referring expression, consequently it is more likely that the coreference exists for the given sentence.

A positive value is again added should the entity and referring expression both exist in the indexed sentence. Another positive value is added every time the referring expression exists following the entity's introduction, so long as there is no introduction of a new entity. Should a new entity be introduced, a positive value will be added to both entities at the given referring expression for the sentence, and the next instance of the referring expression will only increase in value for the pairing of itself and the newly introduced entity. This ensures that should the entity change, the referring expression is equally likely to refer to either entity. However, in the next instance, it is much more likely to refer to the newly introduced entity, and should be accounted for. Penalization does not occur in this instance. The values, before smoothing, are represented as:

Referring Expression x Main Entity	Sent1	Sent 2	Sent 3	Sent 4
It-Basketball Court	0	0	0	2
His-Bob	0	0	3	0
It-Building	0	0	0	3
Her-Mary	4	0	0	0
It- Skating Rink	0	0	0	4

And after the application of smoothing via the chi-squared test, $X^2 = \sum(Y-Z)^2/(Z)$, the matrix is represented by:

Referring Expression x Main Entity	Sent1	Sent 2	Sent 3	Sent 4
It-Basketball Court	0	nan	0	2.25
His-Bob	0	nan	2.53	0
It-Building	0	nan	0	7.6
Her-Mary	8	nan	0	0

It- Skating Rink	0	nan	0	18.0

The highest of the values is then obtained to answer the questions, "who is __" where the blank space is left open for the insertion of the referring expression.

IV. Results.

There are several, notable variables that have caused the program to be, overall unsuccessful. To define success, the approach was meant to design a program that created a significant pairing between the main entities corefered to and their identical referring expression, as well as one that could differentiate between two entities of ambiguous gender and match them with two referring expressions of a given gender. Unfortunately, the data set is too small to account for the second instance. Instead, the programming disregards the ambiguous coreference. As for matching a single referring expression to multiple entities, it seems the program was somewhat successful. While not entirely accurate in identifying that the three entities were indeed referring to the same referring expression, each entity was given a positive value towards being the accurate match to the referring expression. In fact, due to the closeness of value before smoothing, it is quite possible that the smoothing technique used to test for significance may have been the underlying root of inaccurate calculations, or perhaps just not fitted for the data. It is fully plausible that another form of testing for significance will provide better results and show even closer values between the three coreference pairs. Should this be the case, it is a matter of determining via SVM what the hyperplane and room for error would be. Given these results, there is too much distance between the pairs for there to be any statistical significance.

In addition, a larger, more diverse corpus to be trained and tested on would prove to be highly functional, as opposed to one so small. By increasing the corpus, the value for each entity, referring expression, and pair of the two becomes increasingly more accurate in calculations.

However, it is important to take into account of what a solely syntactic evaluation of coreferencing would look like. The idea behind the approach was to attempt to create a semantic element and apply it to a syntactic process to help support the idea of a lack of unification between semantic and syntactic processing of coreference resolution. Because the purpose of the approach was to create a program that could help to increase the accuracy of coreference resolution using semantics, it is still plausible that this approach did succeed, as the accuracy of the coreferencing is higher after the application of this program than with a simple linear evaluation, and takes into account several features, including gender, entity changes, and temporal expressions.

In the future, it may be helpful to evaluate what exactly triggers a temporal switch. For example, in the instance of a non-existing entity and referring expression, it may be helpful to incorporate a data-base of all known entities in the given reality, and create a feature and value based on the existence, or lack thereof. If the entity doesn't exist, it's even likely to completely remove the entity and referring expression pair, such as done in this paper's approach with "she-building", "she-Mary", and "him-Bob".

V. Conclusion.

While the central goal of the approach was not obtained, it has created a potentially useful form of assistance in coreference resolution. In order to improve the quality of the approach, there are multiple actions to take. Firstly, creating not only a POS tagger and parser, but also using a larger, more broadly defined corpus, assumedly one consisting of many wh-questions. Secondly, it would be largely beneficial to have an accessible database that revealed whether a particular entity was in existence, to account for implausible assumptions in real-world situations. It would also be greatly beneficial, if neither of the above are of much assistance, for an additional take on the smoothing and significance testing for each element in the layers, as the chi squared test does not seem appropriate for the results given. This includes using SVM to better define the cutoff point between several entities coreferenced by referring the same expressions, calling for boundaries on the linear regression used to calculate the weight and significance per feature.

Naturally, accounting for further features

will improve the chances of accuracy, and it is encouraged that such an idea be implemented. To increase the likelihood of a highly accurate approach. Since this topic is not one that is commonly explored, this may be grounds for some additional assistance towards optimization of coreference resolution in wh-movement sentences. Overall, the approach taken in this paper is not optimal, but with the right changes, can be implemented for further use.

VI. References.

Anderson, A., Garrod, S. C., & Sanford, A. J. 1983. The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *Quarterly Journal of Experimental Psychology*. *35*(3): 427-440.

Bishop, Christopher M. 2009. "Pattern Recognition and Machine Learning" 8th edition. Springer Publishing.

Duffy SA, Henderson JM, Morris RK. 1989. Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology*. 15:791–801.

Elman JL. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*. 33:1–36.

Galvas, Goron & Snajder, Jan. 2013. Exploring Coreference Uncertainty of Generically Extracted Event Mentions. *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30. 1:*7816.

Garrod, S., & Sanford, T. 1988. Thematic subjecthood and cognitive constraints on discourse structure. *Journal of Pragmatics*. *12*(5): 519-534.

Garrod S, Sanford AJ. 1994. Resolving sentences in a discourse context: How discourse representation affects language understanding. In: Gernsbacher M, editor. *Handbook of psycholinguistics.* New York: Academic Press. pp. 675–698.

Gordon P, Hendrick R. 1998. The representation and processing of coreference in discourse. *Cognitive Science*. 22:389–424.

Grosz, B. J., Weinstein, S., & Joshi, A. K. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2): 203-225.

Grosz, B. J. 1977. The representation and use of focus in a system for understanding dialogs. In *IJCAI* 67:76.

Hess DJ, Foss DJ, Carroll P. 1995. Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology*. 124:62–82.

Hobbs, Jerry R., 1976. ``Pronoun Resolution". *Research Report*. 76-1. University of New York.

Huang, Y. T., & Gordon, P. C. 2011. Distinguishing the time-course of lexical and discourse processes through context, co-reference, and quantified expressions. *Journal of Experimental Psychology. 37*(4), 966–978.

Joseph H, Liversedge S, Simon P, Blythe H, White S, Gathercole S, Rayner K. 2008. Children's and adults' processing of anomaly and implausibility during reading: Evidence from eye movements. *The Quarterly Journal of Experimental Psychology*. 61:708–723.

Kamp, H. and U. Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Lappin, S., & Leass, H. J. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*. *20*(4), 535-561.

Levy R. 2008. Expectation-based syntactic comprehension. *Cognition*. 106:1126–1177.

MacDonald M, Pearlmutter N, Seidenberg M. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review*. 101:676–703.

Mitkov, R. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th*

Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 2:869-875. Association for Computational Linguistics.

Morris RK. 1994. Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology*. 20:92–103.

Moxey L, Sanford A. 1993. Communicating quantities. Hove, UK: Erlbaum.

Moxey L, Sanford A. 2000. Focus effects with negative quantifiers. In: Crocker M, Pickering M, Clifton C, editors. *Architectures and Mechanisms of Language Processing*. Cambridge: Cambridge University Press.

Poesio, M. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*. 72-79. Association for Computational Linguistics.

Poesio, M., Ishikawa, T., Im Walde, S. S., & Vieira, R. 2002. Acquiring Lexical Knowledge for Anaphora Resolution. In *LREC*.

Rayner K, Warren T, Juhasz B, Liversedge S. 2004. The effect of plausibility on eye-movements in reading. *Journal of Experimental Psychology*. 30:1290–1301.

Seidenherg MS, Waters GS, Sanders M, Langer P. 1984. Pre- and post-lexical loci of contextual effects on word recognition. *Memory and Cognition.* 12:315–328.

Sidner, C. L. 1979. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse* (No. AI-TR-537). MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB.

Tanenhaus MK, Spivey-Knowlton M, Eberhard K, Sedivy J. 1995. Integration of visual and linguistic information in spoken language comprehension. *Cognition.* 268:1632.

van Eijck, J. and H. Kamp. 1997. "Representing discourse in context". In J. van Benthem and A. ter Meulen (eds.), *Handbook of Logic and Language.* Amsterdam: Elsevier Science, 179-237.

Warren T, McConnell K, Rayner K. 2008. Effects of context on eye-movements when reading about possible and impossible events. *Journal of Experimental Psychology.* 34:1001–1010

Wiemer-Hastings, P., & Iacucci, C. 2001. A computational model of human coreference judgements. In *Proceedings of the First Workshop on Cognitively Plausible Models of Semantic Processing (SEMPRO 2001).* University of Edinburgh, Human Communication Research Centre, Edinburgh.