The Life and Death of Discourse Entities: Identifying Singleton Mentions

Marta Recasens
Linguistics Department
Stanford University
Stanford, CA 94305
recasens@google.com

Marie-Catherine de Marneffe
Linguistics Department
The Ohio State University
Columbus, OH 43210
mcdm@ling.osu.edu

Christopher Potts
Linguistics Department
Stanford University
Stanford, CA 94305
cgpotts@stanford.edu

Abstract

A discourse typically involves numerous entities, but few are mentioned more than once. Distinguishing discourse entities that die out after just one mention (singletons) from those that lead longer lives (coreferent) would benefit NLP applications such as coreference resolution, protagonist identification, topic modeling, and discourse coherence. We build a logistic regression model for predicting the singleton/coreferent distinction, drawing on linguistic insights about how discourse entity lifespans are affected by syntactic and semantic features. The model is effective in its own right (78% accuracy), and incorporating it into a state-of-the-art coreference resolution system yields a significant improvement.

1 Introduction

Not all discourse entities are created equal. Some lead long lives and appear in a variety of discourse contexts (**coreferent**), whereas others never escape their birthplaces, dying out after just one mention (**singletons**). The ability to make this distinction based on properties of the NPs used to identify these referents (mentions) would benefit not only coreference resolution, but also topic analysis, textual entailment, and discourse coherence.

The existing literature provides numerous generalizations relevant to answering the question of whether a given discourse entity will be singleton or coreferent. These involve the internal syntax and morphology of the target NP (Prince, 1981a; Prince, 1981b; Wang et al., 2006), the grammatical function and discourse role of that NP (Chafe, 1976; Hobbs, 1979; Walker et al., 1997; Beaver, 2004), and the interaction of all of those features with semantic operators like negation, modals, and attitude predicates (Karttunen, 1973; Karttunen, 1976; Kamp, 1981; Heim, 1982; Heim, 1992; Roberts, 1990; Groenendijk and Stokhof, 1991; Bittner, 2001).

The first step in our analysis is to bring these insights together into a single logistic regression model — the *lifespan model* — and assess their predictive power on real data. We show that the features generally behave as the existing literature leads us to expect, and that the model itself is highly effective at predicting whether a given mention is singleton or coreferent. We then provide an initial assessment of the engineering value of making the singleton/coreferent distinction by incorporating our lifespan model into the Stanford coreference resolution system (Lee et al., 2011). This addition results in a significant improvement on the CoNLL-2012 Shared Task data, across the MUC, B^3, CEAF, and CoNLL scoring algorithms.

2 Data

All the data used throughout the paper come from the CoNLL-2012 Shared Task (Pradhan et al., 2012), which included the 1.6M English words from OntoNotes v5.0 (Hovy et al., 2006) that have been annotated with different layers of annotation (coreference, parse trees, etc.). We used the training, development (dev), and test splits as defined in the shared task (Table 1). Since the OntoNotes coreference annotations do not contain singleton mentions, we automatically marked as singletons all the NPs

Dataset	Docs	Tokens	MENTIONS Coreferent	Singletons
Training	2,802	1.3M	152,828	192,248
Dev	343	160K	18,815	24,170
Test	348	170K	19,392	24,921

Table 1: CoNLL-2012 Shared Task data statistics. We added singletons (NPs not annotated as coreferent).

not annotated as coreferent. Thus, our singletons include non-referential NPs but not verbal mentions.

3 Predicting lifespans

Our lifespan model makes a binary distinction between discourse referents that are not part of a coreference chain (singletons) and items that are part of one (coreferent). The distribution of lifespans in our data (Figure 1) suggests that this is a natural division. The propensity of singletons also highlights the relevance of detecting singletons for a coreference system. We fit a binary logistic regression model in R (R Core Team, 2012) on the training data, coding singletons as "0" and coreferent mentions as "1". Throughout the following tables of coefficient estimates, positive values favor coreferents and negative ones favor singletons. We turn now to describing and motivating the features of this model.

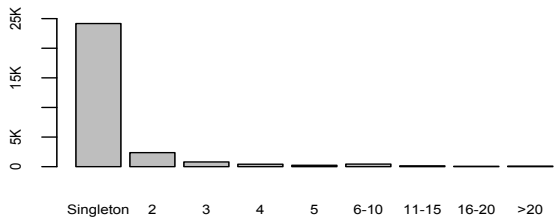

Figure 1: Distribution of lifespans in the dev set. Singletons account for 56% of the data.

Internal morphosyntax of the mention Table 2 summarizes the features from our model that concern the internal morphology and syntactic structure of the mention. Many are common in coreference systems (Recasens and Hovy, 2009), but our model highlights their influence on lifespans. The picture is expected on the taxonomy of given and new defined by Prince (1981b) and assumed throughout dynamic semantics (Kamp, 1981; Heim, 1982): pronouns depend on anaphoric connections to previous mentions for disambiguation and thus are very likely to be coreferent. This is corroborated by the positive coefficient estimate for 'Type = pronoun' in Table 2. Few quantified phrases easily participate in discourse anaphora (Partee, 1987; Wang et al., 2006), accounting for the association between quantifiers and singletons (negative coefficient estimate for 'Quantifier = quantified' in Table 2). The one surprise is the negative coefficient for indefinites. In theories stretching back to Karttunen (1976), indefinites function primarily to establish new discourse entities, and should be able to participate in coreference chains, but here the association with such chains is negative. However, interactions explain this fact (see Table 4 and our discussion of it).

The person, number, and animacy values suggest that singular animates are excellent coreferent NPs, a previous finding of Centering Theory (Grosz et al., 1995; Walker et al., 1998) and of cross-linguistic work on obviative case-marking (Aissen, 1997).

Our model also includes named-entity features for all of the eighteen OntoNotes entity-types (omitted from Table 2 for space and clarity reasons). As a rule, they behave like 'Type = proper noun' in associating with coreferents. The exceptions are ORDINAL, PERCENT, and QUANTITY, which seem intuitively unlikely to participate in coreference chains.

	Estimate	P-value
Type = pronoun	1.21	< 0.001
Type = proper noun	1.88	< 0.001
Animacy = inanimate	-1.36	< 0.001
Animacy = unknown	-0.38	< 0.001
Person = 1	1.05	< 0.001
Person = 2	0.13	< 0.001
Person = 3	1.62	< 0.001
Number = singular	0.61	< 0.001
Number = unknown	0.17	< 0.001
Quantifier = indefinite	-1.49	< 0.001
Quantifier = quantified	-1.23	< 0.001
Number of modifiers	-0.39	< 0.001

Table 2: Internal morphosyntactic features.

Grammatical role of the mention Synthesizing much work in Centering Theory and information structuring, we conclude that coreferent mentions are likely to appear as core verbal arguments and will favor sentence-initial (topic-tracking) positions (Ward and Birner, 2004). The coefficient estimates

	Estimate	P-value
Sentence Position = end	−0.22	< 0.001
Sentence Position = first	0.04	0.07
Sentence Position = last	−0.31	< 0.001
Sentence Position = middle	−0.11	< 0.001
Relation = noun argument	0.56	< 0.001
Relation = other	−0.67	< 0.001
Relation = root	−0.61	< 0.001
Relation = subject	0.65	< 0.001
Relation = verb argument	0.32	< 0.001
In coordination	−0.48	< 0.001

Table 3: Grammatical role features.

in Table 3 corroborate these conclusions. To define the 'Relation' and 'In coordination' features, we used the Stanford dependencies (de Marneffe et al., 2006) on the gold constituents.

Semantic environment of the mention Table 4 highlights the complex interactions between discourse anaphora and semantic operators. These interactions have been a focus of logical semantics since Karttunen (1976), whose guiding observation is semantic: an indefinite interpreted inside the scope of a negation, modal, or attitude predicate is generally unavailable for anaphoric reference outside of the scope of that operator, as in *Kim didn't understand* $[$*an exam question*$]_i$. $^{\#}$*It*$_i$ *was too hard*. Of course, such discourses cohere if the indefinite is interpreted as taking wide scope ('there is a question Kim didn't understand'). Such readings are often disfavored, but they become more salient when modifiers like *certain* are included (Schwarzschild, 2002) or when the determiner is sensitive to the polarity or intensionality of its environment (Baker, 1970; Ladusaw, 1980; van der Wouden, 1997; Israel, 1996; Israel, 2001; Giannakidou, 1999). Subsequent research identified many other factors that further extend or restrict the anaphoric potential of an indefinite (Roberts, 1996).

We do not have direct access to semantic scope, but we expect syntactic scope to correlate strongly with semantic scope, so we used dependency representations to define features capturing syntactic scope for negation, modal auxiliaries, and a broad range of attitude predicates. These features tend to bias in favor of singletons because they so radically restrict the possibilities for intersentential anaphora. Interacting these features with those for the internal syntax of mentions is also informative. Since proper names and pronouns are not scope-taking, they are largely unaffected by the environment features, whereas indefinites emerge as even more restricted, just as Karttunen and others would predict.

Attitude predicates seem initially anomalous, though. They share the relevant semantic properties with negation and modals, and yet they seem to facilitate coreference. Here, the findings of de Marneffe et al. (2012) seem informative. Those authors find that, in texts of the sort we are studying, attitude predicates are used predominantly to mark the source of information that is effectively asserted despite being embedded (Rooryck, 2001; Simons, 2007). That is, though *X said p* does not semantically entail *p*, it is often interpreted as a commitment to *p*, which correspondingly elevates mentions in *p* to main-clause status (Harris and Potts, 2009).

	Estimate	P-value
Presence of negation	−0.18	< 0.001
Presence of modality	−0.22	< 0.001
Under an attitude verb	0.03	0.01
AttitudeVerb * (Type = pronoun)	0.29	< 0.001
AttitudeVerb * (Type = proper noun)	0.14	< 0.001
Modal * (Type = pronoun)	0.12	0.04
Modal * (Type = proper noun)	0.35	< 0.001
Negation * (Type = pronoun)	1.07	< 0.001
Negation * (Type = proper noun)	0.30	< 0.001
Negation * (Quantifier = indefinite)	−0.37	< 0.001
Negation * (Quantifier = quantified)	−0.36	0.23
Negation * (Number of modifiers)	0.11	< 0.001

Table 4: Semantic environment features and interactions.

Results The model successfully learns to tease singletons and coreferent mentions apart. Table 5 summarizes its performance on the dev set. The STANDARD model uses 0.5 as the decision boundary, with 78% accuracy. The CONFIDENT model predicts singleton if Pr $<$ .2 and coreferent if Pr $>$ .8, which increases precision (P) at a cost to recall (R).

	STANDARD			CONFIDENT		
Prediction	R	P	F1	R	P	F1
Singleton	82.3	79.2	80.7	50.5	89.6	64.6
Coreferent	72.2	76.1	74.1	41.3	86.8	55.9

Table 5: Recall, precision, and F1 for the lifespan model.

System	MUC R	MUC P	MUC F1	B^3 R	B^3 P	B^3 F1	CEAF-ϕ_3 R / P / F1	CEAF-ϕ_4 R	CEAF-ϕ_4 P	CEAF-ϕ_4 F1	CoNLL F1
Baseline	**66.64***	64.72	65.67	**68.05***	71.58	**69.77***	58.31	45.49	**47.55***	46.50	60.65
w/ Lifespan	66.08	**67.33***	**66.70***	66.40	**73.14***	69.61	**58.83***	**47.77***	46.38	**47.07***	**61.13***

Table 6: Performance on the test set according to the official CoNLL-2012 scorer. Scores are on automatically predicted mentions. Stars indicate a statistically significant difference (paired Mann-Whitney U-test, $p < 0.05$).

System	B^3 R	B^3 P	B^3 F1	CEAF-ϕ_3 R	CEAF-ϕ_3 P	CEAF-ϕ_3 F1	CoNLL F1
Baseline	**58.53***	71.58	64.40	**63.71***	58.31	60.89	58.86
w/ Lifespan	58.14	**73.14***	**64.78***	63.38	**58.83***	**61.02**	**59.52***

Table 7: B^3, CEAF-ϕ_3 and CoNLL measures on the test set according to a modified CoNLL-2012 scorer that follows Cai and Strube (2010). Scores are on automatically predicted mentions.

4 Application to coreference resolution

To assess the usefulness of the lifespan model in an NLP application, we incorporate it into the Stanford coreference resolution system (Lee et al., 2011), which we take as our baseline. This was the highest-scoring system in the CoNLL-2011 Shared Task, and was also part of the highest-scoring system in the CoNLL-2012 Shared Task (Fernandes et al., 2012). It is a rule-based system that includes a total of ten rules (or "sieves") for entity coreference, such as exact string match and pronominal resolution. The sieves are applied from highest to lowest precision, each rule adding coreference links.

Incorporating the lifespan model The lifespan model can improve coreference resolution in two different ways: (i) mentions classified as singletons should not be considered as either antecedents or coreferent, and (ii) mentions classified as coreferent should be linked with another mention(s). By successfully predicting singletons (i), we can enhance the system's precision; by successfully predicting coreferent mentions (ii), we can improve the system's recall. Here we focus on (i) and use the lifespan model for detecting singletons. This decision is motivated by two factors. First, given the large number of singletons (Figure 1), we are more likely to see a gain in performance from discarding singletons. Second, the multi-sieve nature of the Stanford coreference system does not make it straightforward to decide which antecedent a mention should be linked to even if we know that it is coreferent. We leave the incorporation of coreferent predictions for future work.

To integrate the singleton model into the Stanford coreference system, we let a sieve consider whether a pair of mentions is coreferent only if neither of the two mentions are classified as singletons by our CONFIDENT model. Experiments on the dev set showed that the model often made wrong predictions for NEs. We do not trust the model for NE mentions. Performance on coreference (on the dev set) was higher with the CONFIDENT model than with the STANDARD model.

Results and discussion To evaluate the coreference system with and without the lifespan model, we used the English dev and test sets from the CoNLL-2012 Shared Task, presented in Section 2. Although the CoNLL shared task evaluated systems on only multi-mention (i.e., non-singleton) entities, by stopping singletons from being linked to multi-mention entities, we expected the lifespan model to increase the system's precision. Our evaluation uses five of the measures given by the CoNLL-2012 scorer: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF-ϕ_3 and CEAF-ϕ_4 (Luo, 2005), and the CoNLL official score (Denis and Baldridge, 2009). We do not include BLANC (Recasens and Hovy, 2011) because it assumes gold mentions and so is not suited for the scenario considered in this paper, which uses automatically predicted mentions.

Table 6 summarizes the test set performance. All the scores are on automatically predicted mentions. We use gold POS, parse trees, and NEs. The base-

line is the Stanford system, and 'w/ Lifespan' is the same system extended with our lifespan model to discard singletons, as explained above.

As expected, the lifespan model increases precision but decreases recall. Overall, however, we obtain a significant improvement of 0.5–1 points in the F1 score of MUC, CEAF-ϕ_3, CEAF-ϕ_4 and CoNLL. The drop in B^3 traces to a bug in the CoNLL scorer's implementation of Cai and Strube (2010)'s algorithm for aligning gold and automatically predicted mentions, which affects the computation of B^3 and CEAF-ϕ_3.[1] Table 7 presents the results after modifying the CoNLL-2012 scorer to compute B^3 and CEAF-ϕ_3 according to Cai and Strube (2010).[2] We do see an improvement in the precision and F1 scores of B^3, and the overall CoNLL score remains significant. The CEAF-ϕ_3 F1 score is no longer significant, but is still in the expected direction.

5 Conclusion

We built a model to predict the lifespan of discourse referents, teasing apart singletons from coreferent mentions. The model validates existing linguistic insights and performs well in its own right. This alone has ramifications for tracking topics, identifying protagonists, and modeling coreference and discourse coherence. We applied the lifespan model to coreference resolution, showing how to incorporate it effectively into a state-of-the-art rule-based coreference system. We expect similar improvements with machine-learning-based coreference systems, where incorporating all the power of the lifespan model would be easier.

Our lifespan model has been integrated into the latest version of the Stanford coreference resolution system.[3]

[1] At present, if the system links two mentions that do not exist in the gold standard, the scorer adds two singletons to the gold standard. This results in a higher B^3 F1 score (when it should be lower) because recall increases instead of staying the same (precision goes up).

[2] In the modified scorer, twinless predicted mentions are added to the gold standard to compute precision but not to compute recall.

[3] `http://nlp.stanford.edu/software/dcoref.shtml`

Acknowledgments

We thank Emili Sapena for modifying the CoNLL-2012 scorer to follow Cai and Strube (2010).

This research was supported in part by ONR grant No. N00014-10-1-0109 and ARO grant No. W911NF-07-1-0216. The first author was supported by a Beatriu de Pinós postdoctoral scholarship (2010 BP-A 00149) from Generalitat de Catalunya.

References

Judith Aissen. 1997. On the syntax of obviation. *Language*, 73(4):705–750.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the LREC 1998 Workshop on Linguistic Coreference*, pages 563–566.

C. L. Baker. 1970. Double negatives. *Linguistic Inquiry*, 1(2):169–186.

David Beaver. 2004. The optimization of discourse anaphora. *Linguistics and Philosophy*, 27(1):3–56.

Maria Bittner. 2001. Surface composition as bridging. *Journal of Semantics*, 18(2):127–177.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of SIGDIAL 2010*, pages 28–36.

Wallace L. Chafe. 1976. Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In Charles N. Li, editor, *Subject and Topic*, pages 25–55. Academic Press, New York.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42:87–96.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Proceedings of CoNLL-2012: Shared Task*, pages 41–48.

Anastasia Giannakidou. 1999. Affective dependencies. *Linguistics and Philosophy*, 22(4):367–421.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–100.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Jesse A. Harris and Christopher Potts. 2009. Perspective-shifting with appositives and expressives. *Linguistics and Philosophy*, 32(6):523–552.

Irene Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, UMass Amherst.

Irene Heim. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, 9(2):183–221.

Jerry R. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of HLT-NAACL 2006*, pages 57–60.

Michael Israel. 1996. Polarity sensitivity as lexical semantics. *Linguistics and Philosophy*, 19(6):619–666.

Michael Israel. 2001. Minimizers, maximizers, and the rhetoric of scalar reasoning. *Journal of Semantics*, 18(4):297–331.

Hans Kamp. 1981. A theory of truth and discourse representation. In Jeroen Groenendijk, Theo M. V. Janssen, and Martin Stockhof, editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematical Centre, Amsterdam.

Lauri Karttunen. 1973. Presuppositions and compound sentences. *Linguistic Inquiry*, 4(2):169–193.

Lauri Karttunen. 1976. Discourse referents. In James D. McCawley, editor, *Syntax and Semantics*, volume 7: Notes from the Linguistic Underground, pages 363–385. Academic Press, New York.

William A. Ladusaw. 1980. On the notion 'affective' in the analysis of negative polarity items. *Journal of Linguistic Research*, 1(1):1–16.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. In *Proceedings of CoNLL-2011: Shared Task*, pages 28–34.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP 2005*, pages 25–32.

Barbara H. Partee. 1987. Noun phrase interpretation and type-shifting principles. In Jeroen Groenendijk, Dick de Jong, and Martin Stokhof, editors, *Studies in Discourse Representation Theory and the Theory of Generalized Quantifiers*, pages 115–143. Foris Publications, Dordrecht.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of EMNLP and CoNLL-2012: Shared Task*, pages 1–40.

Ellen Prince. 1981a. On the inferencing of indefinite 'this' NPs. In Bonnie Lynn Webber, Ivan Sag, and Aravind Joshi, editors, *Elements of Discourse Understanding*, pages 231–250. Cambridge University Press, Cambridge.

Ellen F. Prince. 1981b. Toward a taxonomy of given–new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York.

R Core Team, 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Marta Recasens and Eduard Hovy. 2009. A deeper look into features for coreference resolution. In Sobha Lalitha Devi, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications*, volume 5847 of *Lecture Notes in Computer Science*, pages 29–42. Springer.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Craige Roberts. 1990. *Modal Subordination, Anaphora, and Distributivity*. Garland, New York.

Craige Roberts. 1996. Anaphora in intensional contexts. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 215–246. Blackwell Publishers, Oxford.

Johan Rooryck. 2001. Evidentiality, Part II. *Glot International*, 5(5):161–168.

Roger Schwarzschild. 2002. Singleton indefinites. *Journal of Semantics*, 19(3):289–314.

Mandy Simons. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6):1034–1056.

Ton van der Wouden. 1997. *Negative Contexts: Collocation, Polarity and Multiple Negation*. Routledge, London and New York.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of MUC-6*, pages 45–52.

Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors. 1997. *Centering in Discourse*. Oxford University Press.

Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. 1998. Centering in naturally-occurring discourse: An overview. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 1–28, Oxford. Clarendon Press.

Linton Wang, Eric McCready, and Nicholas Asher. 2006. Information dependency in quantificational subordination. In Klaus von Heusinger and Ken Turner, editors,

Where Semantics Meets Pragmatics, pages 267–304. Elsevier Science, Amsterdam.

Gregory Ward and Betty Birner. 2004. Information structure and non-canonical syntax. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, pages 153–174. Blackwell, Oxford.