Cascading Collective Classification for Bridging Anaphora Recognition Using a Rich Linguistic Feature Set

Yufang Hou¹, Katja Markert², Michael Strube¹

¹ Heidelberg Institute for Theoretical Studies gGmbH, Heidelberg, Germany (yufang.hou|michael.strube)@h-its.org ²School of Computing, University of Leeds, UK scskm@leeds.ac.uk

Abstract

Recognizing bridging anaphora is difficult due to the wide variation within the phenomenon, the resulting lack of easily identifiable surface markers and their relative rarity. We develop linguistically motivated discourse structure, lexico-semantic and genericity detection features and integrate these into a cascaded minority preference algorithm that models bridging recognition as a subtask of learning finegrained information status (IS). We substantially improve bridging recognition without impairing performance on other IS classes.

1 Introduction

In *bridging* or *associative anaphora* (Clark, 1975; Prince, 1981; Gundel et al., 1993), the antecedent and anaphor are not coreferent but are linked via a variety of contiguity relations.¹ In Example 1, the phrases *a resident*, *the stairs* and *the lobby* are bridging anaphors with the antecedent *One building*.²

(1) *One building* was upgraded to red status while people were taking things out, and **a resident** called up **the stairs** to his girlfriend, telling her to keep sending things down to **the lobby**.

Bridging is an important problem as it affects linguistic theory and applications alike. For example, without bridging resolution, entity coherence between the first and second coordinated clause in Example 1 cannot be established. This is a problem both for coherence theories such as Centering (Grosz et al., 1995) (where bridging is therefore incorporated as an indirect realization of previous entities) as well as applications relying on entity coherence modelling, such as readability assessment or sentence ordering (Barzilay and Lapata, 2008).

Full bridging resolution needs (i) recognition that a bridging anaphor is present and (ii) identification of the antecedent and contiguity relation. In recent work, these two tasks have been tackled separately, with bridging recognition handled as part of information status (IS) classification (Markert et al., 2012; Cahill and Riester, 2012; Rahman and Ng, 2012). Each mention in a text gets assigned one IS class that describes its accessibility to the reader at a given point in a text, bridging being one possible class. We stay within this framework.

Bridging recognition is a difficult task, so that we had to report very low results on this IS class in previous work (Markert et al., 2012). This is due to the phenomenon's variety, leading to a lack of clear surface features for recognition. Instead, we formulate in this paper novel discourse structure and lexicosemantic features as well as features that distinguish bridging from generics (see Section 3). In addition, making up between 5% and 20% of definite descriptions (Gardent and Manuélian, 2005; Caselli and Prodanof, 2006) and around 6% of all NPs (Markert et al., 2012), bridging is still less frequent than many other IS classes and recognition of minority classes is well known to be more difficult. We therefore use a cascaded classification algorithm to address this problem (Omuya et al., 2013).

¹We exclude *comparative anaphora* where anaphor and antecedent are in a similarity/exclusion relation, indicated by anaphor modifiers such as *other* or *similar* (Modjeska et al., 2003).

²Examples are from OntoNotes (Weischedel et al., 2011). Bridging anaphora are set in boldface; antecedents in italics.

2 Related Work

Most bridging research concentrates on antecedent selection only (Poesio and Vieira, 1998; Poesio et al., 2004a; Markert et al., 2003; Lassalle and Denis, 2011; Hou et al., 2013), assuming that bridging recognition has already been performed. Previous work on recognition is either limited to definite NPs based on heuristics evaluated on small datasets (Hahn et al., 1996; Vieira and Poesio, 2000), or models it as a subtask of learning fine-grained IS (Rahman and Ng, 2012; Markert et al., 2012; Cahill and Riester, 2012). Results within this latter framework for bridging have been mixed: We reported in Markert et al. (2012) low results for bridging in written news text whereas Rahman and Ng (2012) report high results for the four subcategories of bridging annotated in the Switchboard dialogue corpus by Nissim et al. (2004). We believe this discrepancy to be due to differences in corpus size and genre as well as in bridging definition. Bridging in Switchboard includes non-anaphoric, syntactically linked part-of and set-member relationships (such as the building's *lobby*), as well as comparative anaphora, the latter being marked by surface indicators such as other, another etc. Both types are much easier to identify than anaphoric bridging cases.³ In addition, many non-anaphoric lexical cohesion cases have been annotated as bridging in Switchbard as well.

We also separate bridging recognition and antecedent selection. One could argue that a joint model is more attractive as potential antecedents such as *building* "trigger" subsequent bridging cases such as *stairs* (Example 1). However, bridging can be indicated by referential patterns without world knowledge about the anaphor/antecedent NPs, as the nonsense example 2 shows: *the wug* is clearly a bridging anaphor although we do not know the antecedent.⁴

(2) The blicket couldn't be connected to the dax. **The wug** failed.

Similarly, Clark (1975) distinguishes between bridging via necessary, probable and inducible parts/roles and argues that only in the first and maybe the second case the antecedent triggers the bridging anaphor in the sense that we already spontaneously think of the anaphor when we read the antecedent. Also, bridging recognition on its own can be valuable for applications: for example, prosody is influenced by IS status without needing antecedent knowledge (Baumann and Riester, 2013).

3 Characterizing Bridging Anaphora for Automatic Recognition

3.1 Properties of bridging anaphora

Bridging anaphors are rarely marked by surface features. Indeed, even the common practice (Vieira and Poesio, 2000; Lassalle and Denis, 2011; Cahill and Riester, 2012) to limit bridging to definite NPs does not seem to be correct: We report in previous work (Hou et al., 2013) that less than 40% of the bridging anaphora in our corpus are definites. Instead, bridging is diverse with regard to syntactic form and function: bridging anaphora can be definite NPs (Examples 4 and 6), indefinite NPs (Example 5) or bare NPs (Examples 3, 8 and 9). The only frequent syntactic property shared is that bridging NPs tend to have a simple internal structure with regards to modification. Bridging is also easily confused with generics: friends is used as bridging anaphor in Example 9 but generically in Example 10.

(3) ... meat ... The Communists froze prices instead.

(4) ... *the fund's building* ... **The budget** was only \$400,000.

(5) ... *employees* ... A food caterer stashed stones in the false bottom of a milk pail.

(6) ... *his truck* ... The farmer at **the next truck** shouts, "Wheat!"

(7) ... *the firms* ... Crime was the reason that **26%** reported difficulty recruiting personnel and that **19%** said they were considering moving.

(8) ... *the company* ... His father was **chairman** and **chief executive** until his death in an accident five years ago.

(9) ... Josephine Baker ... Friends pitched in.

(10) <u>Friends</u> are part of the glue that holds life and faith together.

Bridging anaphora can have almost limitless variation. However, we observe that bridging anaphors are often licensed because of discourse structure

³See also the high results for our specific category for comparative anaphora (Markert et al., 2012).

⁴We thank an anonymous reviewer for pointing this out.

Markert et al. (2012) local feature set									
f2 FullPreMentionTime (n)									
f4 ContentWordPreMention (b)									
f6 NPtype (n)									
f8 GrammaticalRole (n)									
f10 PreModByCompMarker (b)									
Markert et al. (2012) relational feature set									
f13 Precedes (r)									

Table 1: Markert et al.'s (2012) feature set, b indicates binary, n nominal, r relational features.

and/or lexical or world knowledge. With regard to discourse structure, Grosz et al. (1995) observe that bridging is often needed to establish entity coherence between two adjacent sentences (Examples 1, 2, 4, 5, 6, 7 and 9). With regard to lexical and world knowledge, relational noun phrases (Examples 3, 4, 8 and 9), building parts (Example 1), set membership elements (Example 7), or, more rarely, temporal/spatial modification (Example 6) may favor a bridging reading. Motivated by these observations, we develop discourse structure and lexico-semantic features indicating bridging anaphora as well as features designed to separate genericity from bridging.

3.2 Features

In Markert et al. (2012) we classify eight finegrained IS categories for NPs in written text: *old*, *new* and 6 *mediated* categories (*syntactic*, *world-Knowledge*, *bridging*, *comparative*, *aggregate* and *function*). This feature set (Table 1, f1-f13) works well to identify *old*, *new* and several *mediated* categories. However, it fails to recognize most bridging anaphora which we try to remedy in this work by including more diverse features.

Discourse structure features (Table 2, f1-f3). Bridging occurs frequently in sentences where otherwise there would no entity coherence to previous sentences/clauses (see Grosz et al. (1995) and Poesio et al. (2004b) for discussions about bridging, entity coherence and centering transitions in the Centering framework). This is especially true for *topic* NPs (Halliday and Hasan, 1976) in such sentences.

We follow these insights by identifying coherence gap sentences (see Examples 1, 4, 5, 6, 7, 9 and also 2): a sentence has a coherence gap (f1) if it has none

new local features	s for bridging								
discourse	f1 IsCoherenceGap (b)								
structure	f2 IsSentFirstMention (b)								
	f3 IsDocFirstMention (b)								
semantics	f4 IsWordNetRelationalNoun (b)								
	f5 IsInquirerRoleNoun (b)								
	f6 IsBuildingPart (b)								
	f7 IsSetElement (b)								
	f8 PreModSpatialTemporal (b)								
	f9 IsYear (b)								
	f10 PreModifiedByCountry (b)								
generic	f11 AppearInIfClause (b)								
NP	f12 VerbPosTag (l)								
features	f13 IsFrequentGenericNP (b)								
	f14 WorldKnowledgeNP (l)								
	f15 PreModByGeneralQuantifier (b)								
other features	f16 Unigrams (l)								
0	f17 BridgingHeadNP (1)								
	f18 HasChildNP (b)								
new features for o	other mediated categories								
aggregate	f19 HasChildCoordination (r)								
function	f20 DependOnChangeVerb (b)								
worldKnowledge	f21 IsFrequentProperName (b)								

Table 2: New feature set, l indicates lexical features.

of the following three coherence elements: (1) entity coreference to previous sentences, as approximated via string match or presence of pronouns, (2) comparative anaphora approximated by mentions modified via a small set of comparative markers (see also Table 1, f10 *PreModByCompMarker*), or (3) proper names. We approximate the topic of a sentence via the first mention (f2).

f3 models that bridging anaphors do not appear at the beginning of a text.

Semantic features (Table 2, f4-f10). In contrast to generic patterns, our semantic features capture lexical properties of nouns that make them more likely to be the head of a bridging NP. We create f4-f8 to capture four kinds of bridging anaphora.

Löbner (1985) distinguishes between relational nouns that take on at least one obligatory semantic role (such as *friend*) and sortal nouns. It is likely that relational nouns are more frequently used as bridging than sortal nouns (see Examples 3, 4, 8 and 9). We extract a list containing around 4,000 relational nouns from WordNet and a list containing around 500 nouns that specify professional roles from the General Inquirer lexicon (Stone et al., 1966), then determine whether the NP head appears in these lists or not (f4 and f5). The obligatory semantic role for a relational noun can of course also be filled NP internally instead of anaphorically and we use the features f10 (for instances such as *the Egyptian president*) and f18 (for complex NPs that are likely to fill needed roles NP internally) to address this.

Because part-of relations are typical bridging relations (see Example 1 and Clark (1975)), we use f6 to determine whether the NP is a part of the building or not, using again a list extracted from Inquirer.

f7 is used to identify set membership bridging cases (see Example 7), by checking whether the NP head is a number or indefinite pronoun (such as *none*, *one*, *some*) or modified by *each*, *one*. However, not all numbers are bridging cases (such as 1976) and we use f9 to exclude such cases.

Lassalle and Denis (2011) note that some bridging anaphors are indicated by spatial or temporal modifications (see Example 6). We use f8 to detect this by compiling 20 such adjectives from Inquirer.

Features to detect generic nouns (Table 2, f11f15). Generic NPs (Example 10) are easily confused with bridging anaphora. Inspired by Reiter and Frank (2010) who build on linguistic research, we develop features (f11-f15) to exclude generics.

First, hypothetical entities are likely to refer to generic entities (Mitchell et al., 2002), We approximate this by determining whether the NP appears in an if-clause (f11). Also the clause tense and mood may play a role to decide genericity (Reiter and Frank, 2010). This is often reflected by the main verb of a clause, so we extract its POS tag (f12).

Some NPs are commonly used generically, such as *children*, *men*, or *the dollar*. The ACE-2 corpus (distinct from our corpus) contains generic annotation . We collect all NPs from ACE-2 that are always used generically (f13). We also try to learn NPs that are uniquely identifiable without further description or anaphoric links such as *the sun* or *the pope*. We do this by extracting common nouns which are annotated as *worldKnowledge* from the training part of our corpus⁵ and use these as lexical features (f14).

Finally, motivated by the ACE-2 annotation guidelines, we identify six quantifiers that may indicate genericity, such as *all*, *no*, *neither* (f15).

Other features for bridging (Table 2, f16-f18). Following Rahman and Ng (2012), we use unigrams (f16). We also extract heads of bridging anaphors from the training data as lexical features (f17) to learn typical nouns used for bridging that we did not cover in lexicon extraction (f4 to f6).

Feature f18 models that bridging anaphora most often have a simple internal structure and usually do not contain any other NPs.

Features for other IS categories (Table 2, f19f21). We propose three features to improve other IS categories. In the relational feature f19, we separate coordination parent-child from other parentchild relations to help with the class *aggregate*. f20determines whether a number is the object of an increase/decrease verb (using a list extracted from Inquirer) and therefore likely to be the IS class *function*. Frequent proper names are more likely to be hearer old and hence of the class *worldKnowledge*. f21 extracts proper names that occur in at least 100 documents in the Tipster corpus to approximate this.

4 Experiments and Results

Experimental setup. We perform experiments on the corpus provided in Markert et al. $(2012)^6$. It consists of 50 texts taken from the WSJ portion of the OntoNotes corpus (Weischedel et al., 2011) with almost 11,000 NPs annotated for information status including 663 bridging NPs and their antecedents. All experiments are performed via 10-fold cross-validation on documents. We use gold standard mentions and the OntoNotes named entity and syntactic annotation layers for feature extraction.

Reimplemented baseline system (*rbls*). *rbls* uses the same features as Markert et al. (2012) (Table 1) but replaces the local decision tree classifier with LibSVM as we will need to include lexical features.

rbls + **Table 2 feature set** (*rbls*+*newfeat*). Based on *rbls*, all the new features from Table 2 are added.

Cascading minority preference system (*cmps*). Minority classes such as bridging suffer during standard multi-class classification. Inspired by Omuya

⁵This list varies for each run of our algorithm in 10-fold cross validation.

⁶http://www.h-its.org/nlp/download/ isnotes.php

	collective									cascade + collective					
	markert 12			rbls			rbls+newfeat			cmps			cmps-newfeat		
	R	Р	F	R	Р	F	R	Р	F	R	Р	F	R	Р	F
old	84.1	85.2	84.6	84.6	85.5	85.1	84.4	86.0	85.2	82.2	87.2	84.7	78.9	89.5	83.8
med/worldKnowledge	60.6	70.0	65.0	65.9	69.6	67.7	67.4	77.3	72.0	67.2	77.2	71.9	67.5	66.7	67.1
med/syntactic	75.7	80.1	77.9	77.8	81.2	79.4	82.2	81.9	82.0	81.6	82.5	82.0	73.9	81.7	77.6
med/aggregate	43.1	55.8	48.7	47.9	58.0	52.5	64.5	79.5	71.2	63.5	77.9	70.0	46.9	60.0	52.7
med/function	35.4	53.5	48.7	33.8	56.4	42.3	67.7	72.1	69.8	67.7	72.1	69.8	41.5	50.0	45.4
med/comparative	81.4	82.0	81.7	81.8	82.5	82.1	81.8	82.1	82.0	86.6	78.2	82.2	86.2	78.7	82.3
med/bridging	12.2	41.7	18.9	10.7	36.6	16.6	19.3	39.0	25.8	44.9	39.8	42.2	31.8	23.9	27.3
new	87.7	73.3	79.8	87.5	74.8	80.7	86.5	76.1	81.0	83.0	78.1	80.5	82.4	76.1	79.1
acc		76.8			77.6			78.9			78.6			75.0	

Table 3: Experimental results

et al. (2013), we develop a cascading minority preference system for fine-grained IS classification. For the five minority classes (*function, aggregate, comparative, bridging* and *worldKnowledge*) that each make up less than the expected $\frac{1}{8}$ of the data set, we develop five binary classifiers with LibSVM⁷ using all features from Tables 1 and 2 and apply them in order from rarest to more frequent category. Whenever a minority classifier predicts true, this class is assigned. When all minority classifiers say false, we back off to the multiclass *rbls+newfeat* system.

cmps – **Table 2 feature set** (*cmps*–*newfeat*). To test the effect of using the minority preference system without additional features, we employ a *cmps* system with baseline features from Table 1 only.

Results and Discussion (Table 3). Our novel features in *rbls+newfeat* show improvements for *worldKnowledge, aggregate* and *function* as well as *bridging* categories compared to both baseline systems, although the performance for *bridging* is still low. In addition, the overall accuracy is significantly better than the two baseline systems (at the level of 1% using McNemar's test). Using the cascaded minority preference system *cmps* in addition improves bridging results substantially while the performance on other categories does not worsen. The algorithm needs both our novel feature classes as well as cascaded modelling to achieve this improvement as the comparison to *cmps-newfeat* shows: the latter lowers overall accuracy as it tends to overgenerate rare

classes (including bridging) with low precision if the features are not strong enough. Our novel features (addressing linguistic properties of bridging) and the cascaded algorithm (addressing data sparseness) appear to be complementary.

To look at the impact of features in our best system, we performed an ablation study. Lexical features as well as semantic ones have the most impact. Discourse structure and genericity information features have less of an impact. We believe the latter to be due to noise involved in extracting these features (such as approximating coreference for the coherence gap feature) as well as genericity recognition still being in its infancy (Reiter and Frank, 2010).

5 Conclusions

This paper aims to recognize bridging anaphora in written text. We develop *discourse structure*, *lexico-semantic* and *genericity* features based on linguistic intuition and corpus research. By using a cascading minority preference system, we show that our approach outperforms the bridging recognition in Markert et al. (2012) by a large margin without impairing the performance on other IS classes.

Acknowledgements. Yufang Hou is funded by a PhD scholarship from the Research Training Group *Coherence in Language Processing* at Heidelberg University. Katja Markert receives a Fellowship for Experienced Researchers by the Alexander-von-Humboldt Foundation. We thank HITS gGmbH for hosting Katja Markert and funding the annotation.

⁷Parameter against data imbalance is set according to the ratio between positive and negative instances in the training set.

References

- Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Stefan Baumann and Arndt Riester. 2013. Coreference, lexical givenness and prosody in German. *Lingua*. Accepted.
- Aoife Cahill and Arndt Riester. 2012. Automatically acquiring fine-grained information status distinctions in German. In Proceedings of the SIGdial 2012 Conference: The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Seoul, Korea, 5–6 July 2012, pages 232–236.
- Tommaso Caselli and Irina Prodanof. 2006. Annotating bridging anaphors in Italian: In search of reliability. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 22–28 May 2006.
- Herbert H. Clark. 1975. Bridging. In Proceedings of the Conference on Theoretical Issues in Natural Language Processing, Cambridge, Mass., June 1975, pages 169– 174.
- Claire Gardent and Hélène Manuélian. 2005. Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1):115–140.
- 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.
- Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.
- Udo Hahn, Michael Strube, and Katja Markert. 1996. Bridging textual ellipses. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, 5–9 August 1996, volume 1, pages 496–501.
- M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. London, U.K.: Longman.
- Yufang Hou, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, 9–14 June 2013, pages 907–917.
- Emmanuel Lassalle and Pascal Denis. 2011. Leveraging different meronym discovery methods for bridging resolution in French. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium* (*DAARC 2011*), Faro, Algarve, Portugal, 6–7 October 2011, pages 35–46.

- Sebastian Löbner. 1985. Definites. Journal of Semantics, 4:279–326.
- Katja Markert, Malvina Nissim, and Natalia N. Modjeska. 2003. Using the web for nominal anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*. Budapest, Hungary, 14 April 2003, pages 39–46.
- Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, 8–14 July 2012, pages 795–804.
- Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstain, Lisa Ferro, and Beth Sundheim. 2002. ACE-2 Version 1.0. LDC2003T11, Philadelphia, Penn.: Linguistic Data Consortium.
- Natalia M. Modjeska, Katja Markert, and Malvina Nissim. 2003. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, 11–12 July 2003, pages 176–183.
- Malvina Nissim, Shipara Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 26–28 May 2004, pages 1023–1026.
- Malvina Nissim. 2006. Learning information status of discourse entities. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006, pages 94–012.
- Adinoyi Omuya, Vinodkumar Prabhakaran, and Owen Rambow. 2013. Improving the quality of minority class identification in dialog act tagging. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, 9–14 June 2013, pages 802–807.
- Massimo Poesio and Renata Vieira. 1998. A corpusbased investigation of definite description use. *Computational Linguistics*, 24(2):183–216.
- Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004a. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 21–26 July 2004, pages 143–150.
- Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004b. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3). 309-363.

- Ellen F. Prince. 1981. Towards a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York, N.Y.
- Altaf Rahman and Vincent Ng. 2012. Learning the finegrained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 23–27 April 2012, pages 798– 807.
- Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010, pages 40–49.
- Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and Cambridge Computer Associates. 1966. *General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, Mass.
- Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2011. OntoNotes release 4.0. LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium.