The (Non)Utility of Predicate-Argument Frequencies for Pronoun Interpretation

Andrew Kehler[*]
UC San Diego
akehler@ucsd.edu

Douglas Appelt
SRI International
appelt@ai.sri.com

Lara Taylor[*]
UC San Diego
lmtaylor@ucsd.edu

Aleksandr Simma[†]
UC San Diego
asimma@ucsd.edu

Abstract

State-of-the-art pronoun interpretation systems rely predominantly on morphosyntactic contextual features. While the use of deep knowledge and inference to improve these models would appear technically infeasible, previous work has suggested that predicate-argument statistics mined from naturally-occurring data could provide a useful approximation to such knowledge. We test this idea in several system configurations, and conclude from our results and subsequent error analysis that such statistics offer little or no predictive information above that provided by morphosyntax.

1 Introduction

The last several years has seen a number of works that use weight-based systems (trained either manually or via supervised learning) for pronoun interpretation, in addition to others that have addressed the broader task of entity-level coreference (see Mitkov (2002) for a useful survey). These systems typically rely on a variety of morphosyntactic factors that have been posited in the literature to affect the interpretation of pronouns in naturally-occurring discourse, including gender and number agreement, the distance between the pronoun and antecedent, the grammatical positions of the pronoun and antecedent, and the linguistic form of the antecedent, among others. A common refrain is that the performance of systems that rely on such features is plateauing, and that further progress will require the use of world knowledge and inference (ibid., Ch. 9, inter alia). World knowledge, after all, would seem to play a role in determining that the referent of *it* in example (1) is the entity denoted by *his industry* rather than *Glendening's initiative* or *the edge.*

(1) He worries that Glendening's initiative could push his industry over the edge, forcing it to shift operations elsewhere.

Of course, no well-suited knowledge base and accompanying inference procedure exists that can deliver such a capability robustly in an open domain.

In lieu of this capability, previous authors have suggested that what can be viewed as a more superficial form of semantic information – predicate-argument statistics mined from naturally-occurring data – could be used to capture certain selectional regularities. For instance, such statistics might reveal that *forcing_industry* is a more likely verb-object combination in naturally-occurring data than *forcing_initiative* or *forcing_edge*. Assuming that such statistics imply that industries are more likely to be forced in the real world than are initiatives or edges, this information could be taken to establish a preference for *his industry* as the antecedent of *it* in (1). While there will always be cases that require arbitrarily deep knowledge for their interpretation, the empirical question of how far one can go by relying on this sort of selectional information remains.

Our point of departure is the work of Lappin and Leass (1994, henceforth L&L) and Dagan et al. (1995). (See also Dagan and Itai (1990).) L&L demonstrated with a system called RAP that a (manually-tuned) weight-based scheme for integrating pronoun interpretation preferences can achieve high performance on real data, in their case, 86% accuracy on a corpus of computer training manuals.[1] Dagan et al. (1995) then developed a postprocessor based on predicate-argument statistics that was used to override RAP's decision when it failed to express a clear preference between two or more antecedents, which resulted in a modest rise in per-

[*] Department of Linguistics.

[†] Department of Computer Science and Engineering.

[1] Kennedy and Boguraev (1996, henceforth, K&B) adapted L&L's algorithm to rely on far less syntactic analysis (noun phrase identification and rudimentary grammatical role marking), with performance in the 75% range on mixed genres.

formance (2.5%).[2] Because RAP is symbolic, the two systems were necessarily coupled in a black-box manner. They noted, however, that if one had a statistically-driven pronoun interpretation system, co-occurrence information could be modeled alongside morphosyntactic information:

> "A promising direction for future research is the development of an empirically based model for salience criteria analogous to the one that we constructed for lexical preference. The integration of these models using a probabilistic decision procedure will hopefully yield an optimized integrated system for anaphora resolution." (p. 643)

In this work we set out to evaluate Dagan et al.'s proposal. Indeed, the weight-combination scheme of L&L is suggestive of a particular approach to supervised learning – maximum entropy (MaxEnt) – in which such a system of weights is inferred from maximum likelihood counts on annotated data. Using MaxEnt, we trained a system based on an optimized set of morphosyntactic features and augmented it with predicate-argument statistics in two scenarios: (i) one mimicking the Dagan et al. postprocessor, and (ii) one in which the predicate-argument statistics were represented as features alongside the morphosyntactic features. Our results and subsequent error analysis suggest, however, that such statistics offer little or no predictive information above that provided by morphosyntax.

2 Corpora Used

The training and test data sets came from the newspaper and newswire segments of the Automatic Content Extraction (ACE) program corpus. The training data contained 2773 annotated third-person pronouns, and the test data (the February 2002 evaluation set) contained 762 annotated third-person pronouns. The performance statistics on the test data reported here are from the only time an evaluation with this data was performed; progress during development was estimated solely via jackknifing on the training data.

The annotated pronouns included only those that were ACE "markables", i.e., ones that referred to entities of the following types: PERSONS, ORGANIZATIONS, GEOPOLITICALENTITIES (politically defined geographical regions, their governments, or their people), LOCATIONS, and FACILITIES. Thus, there were pronouns in both the development and (presumably) test sets for which there were no annotations. As such, certain problems that real-world systems face, such as non-referential (e.g., 'pleonastic') pronouns and pronouns that refer to eventualities, did not have to be dealt with. (However, these pronouns were possible antecedents to other pronouns, and thus were sometimes mistakenly selected as the correct antecedent.) Thus, our results are not necessarily comparable to those of a system that deals with these difficulties (although previous work varies a fair bit on how their datasets were filtered in this regard). Our main purpose here is to establish a state-of-the-art baseline with which to assess the contribution of predicate-argument frequency information.

3 Learning Algorithms

We implemented three pronoun interpretation systems: a MaxEnt model, a Naive Bayes model, and a version of the Hobbs algorithm as a baseline. Our experimentation was driven predominantly using the MaxEnt system using an n-fold jackknifing paradigm (n was typically three). Naive Bayes was implemented toward the end of the project as a machine learning baseline. Both machine learning algorithms were trained as binary coreference classifiers, that is, the examples provided to them consisted of pairings of a pronoun and a possible antecedent phrase, along with a binary coreference outcome determined from the annotated keys. Thus, for a given pronoun there was one example generated for each possible antecedent phrase. So as to focus learning on only the coreferential phrase that is most likely to have been directly responsible for a given pronominalization, all coreferential phrases except the closest in terms of Hobbs distance (discussed later) were eliminated before training. Because we are ultimately interested in identifying the correct antecedent among a list of possible ones, during testing the antecedent assigned the highest probability was chosen.

These systems received as input the results of SRI's TEXTPRO system, a chunk-style shallow parser capable of recognizing low-level constituents (noun groups, verb groups, etc.). No difficult attachments are attempted, and the results are errorful. There was no human-annotated linguistic information in the input. The systems are described further below.

Maximum Entropy Modeling As previously indicated, the weight-based scheme of L&L suggests MaxEnt modeling (Berger et al., 1996) as a particularly natural choice for a machine learning approach.

[2] The difference amounted to 9 additional correct predictions in a corpus of 360 examples. They express a belief that the improvement is real, but acknowledge that they would need twice as many examples in their corpus to reach statistical significance.

In MaxEnt, the parameters of an exponential model of the following form are estimated:

$$p(y|x) = \frac{e^{\sum_i \lambda_i f_i(x,y)}}{\sum_y e^{\sum_i \lambda_i f_i(x,y)}}$$

The variable y represents the outcome (coreference or not) and x represents the context. There is one value for each feature that predicts coreference behavior, represented by the parameters $\lambda_1, ..., \lambda_n$, which are Lagrange multipliers that constrain the expected value of each feature in the model to be the values found in the distribution of the training data. (The $f_i(x, y)$ are indicator functions which equal 1 when the corresponding feature is present, and 0 otherwise.) The desired values for these parameters are obtained by maximizing the likelihood of the training data with respect to the model.[3] Thus, whereas L&L's RAP system uses an additive system of weights that is trained manually, the MaxEnt system learns a multiplicative system of weights automatically. One can view the MaxEnt system as yielding a probabilistic notion of antecedent salience: The salience value assigned to a potential antecedent of a given pronoun is just the probability that Maxent assigns to the outcome of coreference.

Naive Bayes In Naive Bayes modeling, a Bayesian probability distribution is estimated under a strong assumption: that all of the features are conditionally independent given the target value. Thus given n features x_i with respect to the context x, we have:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} \approx \frac{p(y)\ \prod_{i=1}^{n} p(x_i|y)}{p(x)}$$

The context x is constant for each outcome y, so we only need to find:

$$\underset{y \in \{0,1\}}{\text{argmax}} \quad p(y) \prod_{i=1}^{n} p(x_i|y)$$

For most natural language processing scenarios, including ours, this independence assumption is almost certainly false. Nonetheless, Naive Bayes models seem to work well in practice when used as classifiers. That is, the choice that receives the highest probability (relative to the other choices) is often the correct one even though the actual probabilities the model generates may not be very good. These models have the advantage that they are efficiently trained; only a single pass through the training data is necessary.

Hobbs Algorithm We also implemented a version of Hobbs's (1978) well-known pronoun interpretation algorithm as a baseline, in which no machine learning is involved. His algorithm takes the syntactic representations of the sentences up to and including the current sentence as input, and performs a search for an antecedent noun phrase on these trees. Since our shallow parsing system does not build full syntactic trees for the input, we developed a version that does a simple search through the list of noun groups recognized. In accordance with Hobbs's search procedure, noun groups are searched in the following order: (i) in the current sentence from right-to-left, starting with the first noun group to the left of the pronoun, (ii) in the previous sentence from left-to-right, (iii) in two sentences prior from left-to-right, and (iv) in the current sentence from left-to-right, starting with the first noun group to the right of the pronoun (for cataphora). The first noun group that agrees with the pronoun with respect to number, gender, and person is chosen as the antecedent.

4 Features

Our automatically trained systems employed a set of HARD CONSTRAINTS and SOFT FEATURES. Hard constraints are used to weed out potential antecedents before they are sent to the machine learning algorithm. There are only two such constraints, one based on number agreement and one based on gender agreement. Both are conservative in their application. The soft features are used by the machine learning algorithm. After considerable experimentation we settled on a set of forty such features, not including predicate-argument features that will be described in Section 5. These features fall into five categories, listed here with abbreviations that will be used in the tables given in Section 6:

Gender Agreement (gend): Includes features to test a strict match of gender (e.g., a male pronoun and male antecedent), as well as mere compatibility (e.g., a male pronoun with an antecedent of unknown gender). These features are more liberal than the gender-based hard constraint mentioned above.

Number Agreement (num): Includes features to test a strict match of number (e.g., a singular pronoun and singular antecedent), as well as mere compatibility (e.g., a singular pronoun with an antecedent of unknown number). These features are likewise more liberal than the number-based hard constraint mentioned above.

[3] The results reported here were produced by using the IMPROVED ITERATIVE SCALING algorithm with binary-valued features. We also experimented with real-valued features, with highly similar results on jackknifed data.

Distance (dist): Includes features pertaining to the distance between the pronoun and the potential antecedent. Examples include the number of sentences between them and the "Hobbs distance", that is, the number of noun groups that Hobbs's search algorithm has to skip before the potential antecedent is found (Hobbs, 1978; Ge et al., 1998).

Grammatical Role (pos): Includes features pertaining to the syntactic position of the potential antecedent. Examples include whether the potential antecedent appears to be the subject or object of a verb, and whether the potential antecedent is embedded in a prepositional phrase.

Linguistic Form (lform): Includes features pertaining to the referential form of the potential antecedent, e.g., whether it is a proper name, definite description, indefinite NP, or a pronoun.

The values of these features – computed from our system's errorful shallow constituent parses – comprised the input to the learning algorithms, along with the outcome as indicated by the annotated key.

5 Predicate-Argument Frequencies

With a trained statistical model for pronoun interpretation in hand, we can now consider the use of predicate-argument statistics to improve it. Consider sentence (1) again, repeated as (2).

(2) He worries that Glendening's initiative could push his industry over the edge, forcing it to shift operations elsewhere.

Suppose that our system selects *the edge* as the antecedent of *it* instead of *his industry.* It turns out that in a large corpus of shallowly-parsed data (particularly the newswire subset of the TDT-2 corpus, see below), *industr(y|ies)* appears nine times as the head of the object noun phrase of *force* (in its various number/tense combinations), whereas *edge(s)* never does.[4] So by collecting predicate-argument co-occurrence statistics, one could extract the "knowledge" that industries are (statistically speaking) more likely to be forced than edges are, and possibly use this information to change the prediction of the statistical model.

We utilized three types of predicate-argument statistics in our experiments: subject-verb, verb-object, and possessive-noun. We processed the entire newswire subset of the Topic Detection and Tracking (TDT-2) corpus with TextPro, which resulted in 1,321,072 subject-verb relationships, 1,167,189 verb-object relationships, and 301,477 possessive-noun relationships. Words were categorized by their lemmas when available, and proper names for each of the ACE entity types were classified into respective classes (i.e., proper person names all counted as instances of PROPER_PERSON).

While counts were collected for a broad range of predicate-argument combinations, there were still many combinations that were only seen once or twice, and certainly other possible combinations exist that were not seen at all. The distribution that these statistics yield therefore needed to be smoothed. We took two approaches to smoothing. First, because Dagan et al. used Good-Turing smoothing in their experiments, we did likewise so as to replicate their work as closely as possible. Second, we tried an approach based on the distributional clustering method of Pereira et al. (1993). This method yielded word classes that offered more robust count approximations for their member words. However, both methods yielded similar results when embedded in the larger system, and so we will report on the results of using Good-Turing so as to remain more directly comparable to Dagan et al.

The smoothed predicate-argument statistics were employed in two ways. First, we built a postprocessing filter modeled directly on Dagan et al.'s system. Their implementation made use of two equations. The first computes the frequency with which a candidate head noun C is found with the predicate word A, normalized by the number of times C is found alone, so as to not bias the statistic towards words that are common in isolation:

$$stat(C) = P(tuple(C,A) \mid C) = \frac{freq[tuple(C,A)]}{freq(C)}$$

The second equation then weighs the difference in statistical co-occurrence against the different salience values assigned by the pronoun interpretation module for two competing candidates C_1 and C_2:

$$ln\left(\frac{stat(C_2)}{stat(C_1)}\right) > K \times [salience(C_1) - salience(C_2)]$$

The parameter K determines the threshold at which statistical preferences supersede salience preferences. In our implementation, the measure of salience is simply the probability of coreference assigned by the statistical model. Another parameter *max* sets a threshold for the maximum difference between the salience values for the two candidates; any pair for

[4] Likewise, the subject-verb combination *industr(y|ies) shift* occurs three times in the corpus whereas *edge(s) shift* does not.

which this difference exceeds *max* will not be considered. For each combination of feature sets that we evaluated (see Section 6), we performed additional experiments to determine the optimal values of K and *max*. To keep consistent with Dagan et al., statistics were not used (here or in the other MaxEnt system) for potential antecedents that were themselves pronominal. To properly use statistics in such cases, the system would need to have access to an antecedent for the pronoun that has a lexical head; neither model was given access to such information.

In our second approach, we simply developed features that represent the magnitude of the predicate-argument statistics and utilized them during MaxEnt training along with the morphosyntactic features described earlier. The statistics were normalized by dividing them by the total counts for the head of the potential antecedent in the relevant predicate-argument configuration.[5]

In certain respects these different system configurations mirror questions about pronoun interpretation that linger in the theoretical and psycholinguistics literature. A result showing that the postprocessing filter version works best might provide evidence, as has been suggested, that people primarily use morphosyntactic features to resolve pronouns, relying on semantic information only when more than one possibility remains active. A result showing that the integrated version works best might suggest that semantic information is used in concert with morphosyntactic information. Finally, a result showing that neither version improves performance might suggest that morphosyntactic information is the dominant determinant of pronoun interpretation, and/or that any semantic information utilized is not obtained primarily from superficial cues. The results are reported in the next section.

6 Results

Our final MaxEnt system used 40 features, which were categorized into five classes in Section 4. To get a sense for the relative contributions of each feature type, we ran evaluations with all 2^5 (32) possible combinations of these five groups. We first report results on the held-out training data, and then provide the blind test results. Table 1 provides the results on the held-out sections of the training data during 3-fold jackknifing for a sample of five of these 32 combinations. The four rightmost columns represent the results from: (i) MaxEnt with no frequency features (MaxEnt), (ii) MaxEnt with frequency features included during training (MaxEnt-Features), (iii) MaxEnt with Dagan et al. postprocessing (MaxEnt-Postprocessing), and (iv) Naive Bayes without frequencies. Experiments with n-fold jackknifing for other values of n produced similar results. Dagan et al. postprocessing was not attempted with Naive Bayes since the postprocessor makes crucial use of the probabilities the model assigns to competing antecedents, and as previously mentioned, the actual probabilities assigned by Naive Bayes are not necessarily reliable due to the independence assumptions it makes.

The testing phase breaks ties with respect to the order imposed by Hobbs's algorithm. In the case in which no features were used during "training" (see the first row of Table 1, columns 2 and 5), the models will produce the same probability for each possible antecedent. Thus, these experiments reduce to using the Hobbs algorithm, which, performing at 68.23% accuracy,[6] provides a nontrivial baseline. As can be seen, adding groups of additional features incrementally improves performance, up to a final result of 76.16% for MaxEnt using all morphosyntactic features, and a comparable 76.24% for Naive Bayes.

In the end, the predicate-argument statistics provided little if any value, used either as features during MaxEnt training or for Dagan et al. postprocessing. In the best-performing MaxEnt system configuration (see bottom row), the statistics improve performance by less than 0.5%. Interestingly, performance was hurt when *only* statistical features were used (65.71% in MaxEnt-Features and 66.25% in MaxEnt-Postprocessing) as compared to none at all (68.23%). Whereas the Hobbs algorithm ranks *all* of the potential antecedents when no features are used, it only breaks ties in the MaxEnt-Features system that remain after statistical features order the potential antecedents, and the MaxEnt-Postprocessor system uses statistics to rerank the Hobbs ordering between potential antecedents after the fact. This reranking proved detrimental in both cases.

Table 2 provides the final results of blind test evaluation for the same five combinations of feature sets. The final result of the system without predicate-argument statistics was 75.72%, which is presumably reasonable performance considering that the system does not rely on fully-parsed input and lacks access

[5] Experiments with unnormalized counts were also run on jackknifed data with similar results.

[6] All results are reported here in terms of accuracy, that is, the number of pronouns correctly resolved divided by the total number of pronouns read in. Correctness is defined with respect to anaphor-antecedent relationships: a chosen antecedent is correct if the ACE keys place the pronoun and antecedent in the same coreference class.

Features	MaxEnt	MaxEnt-Features	MaxEnt-Postprocessing	Naive Bayes
none	.6823	.6571	.6625	.6823
num, gend	.6870	.6863	.6841	.6859
num, gend, dist	.7274	.7386	.7461	.7313
num, gend, dist, pos	.7425	.7465	.7505	.7436
num, gend, dist, pos, lform	.7616	.7663	.7656	.7624

Table 1: Results from jackknifing on training data

to world knowledge.[7] In this case, the integrated feature system performed identically, whereas the postprocessor system displayed a performance improvement of about 1% (a difference of 8 pronouns).

The MaxEnt results on the test data suffered only a minimal (and in a few cases, no) loss from those on the held-out data. Overtraining appears to have been kept to a minimum; the generality of the features was perhaps responsible for this.[8] The results from Naive Bayes generalized less well, exhibiting a 2% decrement on the test evaluation. The Hobbs algorithm, which is not trained, exhibited similar performance on both sets of data.

7 Error Analysis

There are a variety of possible reasons why the predicate-argument statistics failed to markedly improve performance in each of the system configurations. While it could be that such statistics are simply not good predictors for pronoun interpretation, data sparsity in the collected predicate-argument statistics could also be to blame.

We carried out an error analysis to gain further insight into this question. To address the data-sparsity issue, we employed the technique used in Keller and Lapata (2003, K&L) to get a more robust approximation of predicate-argument counts.[9] We wrote a script to collect the number of pages that the AltaVista search engine found for each predicate-argument combination and its variants per the following schema, modeled directly after K&L:

Subject-Verb: Search for occurrences of the combinations N V where N is the singular or plural form of the subject head noun and V is the infinitive, singular or plural present, past, perfect, or gerund of the head verb.

Verb-Object: Search for occurrences of the combinations V Det N, where V and N are as above for the verb and object head noun respectively, and Det is the determiner *the, a(n),* or the empty string.

Possessive-Noun: Search for occurrences of the combinations $Poss$ N, where $Poss$ is the singular or plural form of the possessive and N is the singular or plural form of the noun.

As in K&L, all searches were done as exact matches. The results for all of the different form combinations totaled together comprised the *unnormalized* counts. We also computed *normalized* counts, in which the unnormalized count was divided by the total number of pages AltaVista returned for the head of the candidate antecedent, so that, as before, the counts would not unduly bias antecedents with head words that occurred frequently in isolation.

We created a list of those examples for which the MaxEnt model – trained with all 5 groups of the morphosyntactic features activated, but not any statistical ones – made incorrect predictions during 3-fold jackknifing on the training data. (We used held-out data so that our test data would remain blind.) We then pared the list down to a reasonable size for manual analysis in a variety of ways. First, of course, only those examples that fall into one of the three predicate-argument configurations with which we are concerned were included (most were). Second, we filtered out the cases in which either the most proximal correct antecedent (with proximity defined

[7] These performance results include 64 "impossible" cases in which, due to misparsing, no correct antecedents were provided to the model; hence 91.6% accuracy is the best that could be achieved. The results likewise include errors in which the model selected a bogus antecedent that resulted from a misparse.

[8] As such, informal post-hoc experiments with Gaussian smoothing (Chen and Rosenfeld, 2000) failed to improve performance.

[9] K&L use this technique to obtain frequencies for predicate-argument bigrams that were unseen in a given corpus, showing that the massive size of the web outweighs the noisy and unbalanced nature of searches performed on it to produce statistics that correlate well with corpus data. We are admittedly extending this reasoning to relations between the *heads* of predicates and arguments without establishing that K&L's technique so generalizes, but we nonetheless feel that it is sufficient for the purpose of an exploratory error analysis. The results we received from this technique held few surprises.

Features	MaxEnt	MaxEnt-Features	MaxEnt-Postprocessing	Naive Bayes
none	.6877	.6496	.6627	.6877
num, gend	.6667	.6745	.6719	.6654
num, gend, dist	.7336	.7415	.7428	.7297
num, gend, dist, pos	.7441	.7507	.7520	.7441
num, gend, dist, pos, lform	.7572	.7572	.7677	.7415

Table 2: Results of final blind test evaluation

with respect to the Hobbs algorithm's search order) or the antecedent chosen by the model was a proper name. Because all proper names of each ACE type were classed together in our experiments, statistics would not make different predictions for two such names. While statistics could differentiate between a potential proper name antecedent and one headed by a common noun (and presumably did in our experiments), we could not use K&L's method on those cases unless we used the actual proper name instead of the category in the search query – this would likely create an undue bias to the other antecedent; consider comparing counts for *lawyer argued* with those for *Snodgrass argued*. Third, we filtered out cases in which either the chosen antecedent or most proximal correct antecedent was itself a pronominal, for the reasons given in Section 5. Lastly, we eliminated a small set of cases in which the chosen antecedent was headless, as no predicate-argument statistics could be collected for such a case.

These filters pared down the errors to a corpus of 45 examples; in all cases the chosen antecedent and most proximal correct antecedent were each headed by common nouns. Upon manual inspection, a further subset of the cases were found to be caused by factors irrelevant to the question at hand: 9 cases in which the antecedent chosen should have been ruled out as impossible (e.g., the collocation *these days* as the antecedent of *they*), 5 cases in which either the annotated keys were incorrect or our mapping system failed to assign credit for a correct answer where it was due, and 11 cases in which our shallow parser misparsed either the chosen antecedent or the correct antecedent. This left a corpus of 20 cases to examine using the K&L methodology.

The preferences embodied by the statistics collected split these cases down the middle: in 10 cases the correct antecedent had a higher normalized probability than the chosen one, and in the other 10 cases the opposite was true.[10] To get a sense for the data, we consider two examples, the first being a case in which predicate-argument statistics were definitive:

(3) After the endowment was publicly excoriated for having the temerity to award some of its money to art that addressed changing views of gender and race, many institutions lost the will to show any art that was rambunctious or edgy.

The MaxEnt model selected *the temerity* as the antecedent of *its* (salience value: 0.30), preferring it to the correct antecedent *the endowment* (salience value: 0.10). However, AltaVista found no occurrences of *temerity's money* or its variants on the web, and thus the unnormalized and normalized counts were 0. On the other hand, *endowment's money* and its variants had unnormalized and normalized statistics of 1583 and 1.47×10^{-3} respectively.

Example (4), on the other hand, is a case in which the statistics merely strengthened the bias to the wrong antecedent:

(4) The dancers were joined by about 70 supporters as they marched around a fountain not far from the mayor's office, chanting: "Giuliani _ scared of sex! Who's he going to censor next?"

The model preferred *the supporters* as the antecedent of *they* (salience value: 0.54) over the correct antecedent *the dancers* (salience value: 0.45). Statistics support the same conclusion, with unnormalized and normalized counts of 2283 and 1.18×10^{-3} for *supporters marched* and its variants, and of 334 and 1.72×10^{-4} for *dancers marched* and its variants.

The analysis of this sample therefore suggests that predicate-argument statistics are unlikely to be of much help when used in a model trained with a state-of-the-art set of morphosyntactic features, even if robust counts were available. While the statistical preferences for our data sample were split down the middle, it is important to understand that the cases for which statistics hurt are potentially more damning than those for which they helped. In the cases in which statistics reinforced a wrong answer, no (reasonable) manipulation of statistical features or filters can rescue the prediction. On the other hand, for

[10] The unnormalized counts disagreed with the normalized ones in only one case; the unnormalized one favored the correct antecedent for that example.

the cases in which statistics could help, their successful use will depend on the existence of a formula that can capture these cases without changing the predictions for examples that the model currently classifies correctly. Although our informal analysis admittedly has certain limits – the web counts we collected are only approximations of true counts, and the size of our manually-inspected corpus ended up being fairly small – our experience leads us to believe that predicate-argument statistics are a poor substitute for world knowledge, and more to the point, they do not offer much predictive power to a state-of-the-art morphosyntactically-driven pronoun interpretation system. Indeed, crisp "textbook" examples such as (3) appear to be empirically rare; the help provided by statistics for several of the examples seemed to be more due to fortuity than the capturing of an actual world knowledge relationship. Consider (5):

(5) Chung, as part of a plea bargain deal with the department, has claimed that then-DNC finance director Richard Sullivan personally asked him for a $125,000 donation in April 1995, the sources said. Sullivan took the money despite having previously voiced suspicions that Chung was acting as a conduit for illegal contributions from Chinese business executives, they added.

Our system selected *Chinese business executives* as the antecedent of *they* (salience value: 0.34), over the correct *the sources* (salience value: 0.10). Predicate-argument statistics support *sources* (normalized and unnormalized values of 29662 and 5.78×10^{-4}) over *executives* (2391 and 1.50×10^{-4}), and thus this example was classified as one of the 10 for which statistics helped. In actuality, however, the correct antecedent is determined by unrelated factors, demonstrated by the fact that if the head nouns *executives* and *sources* were switched in (5), the preferred antecedent would be *the executives*, contrary to what predicate-argument statistics would predict.

8 Conclusion

In conclusion, our experimental results and error analysis suggest that predicate-argument statistics offer little predictive power to a pronoun interpretation system trained on a state-of-the-art set of morphosyntactic features. On the one hand, it appears that the distribution of pronouns in discourse allows for a system to correctly resolve a majority of them using only morphosyntactic cues. On the other hand, predicate-argument statistics appear to provide a poor substitute for the world knowledge that may be necessary to correctly interpret the remaining cases.

Acknowledgments

This work was supported by the ACE program (www.nist.gov/speech/tests/ACE/).

References

Adam Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Stanley F. Chen and Ronard Rosenfeld. 2000. A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, 8(1):37–50.

Ido Dagan and Alon Itai. 1990. Automatic acquisition of constraints for the resolution of anaphora references and syntactic ambiguities. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, pages 330–332.

Ido Dagan, John Justenson, Shalom Lappin, Herbert Leass, and Amnon Ribak. 1995. Syntax and lexical statistics in anaphora resolution. *Applied Artificial Intelligence*, 9(6):633–644, Nov/Dec.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Quebec.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3).

Christopher Kennedy and Branimir Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, pages 183–190.