A Pronoun Anaphora Resolution System based on Factorial Hidden Markov Models

Dingcheng Li
University of Minnesota,
Twin Cities, Minnesosta
lixxx345@umn.edu

Tim Miller
University of Wisconsin
Milwaukee, Wisconsin
tmill@cs.umn.edu

William Schuler
The Ohio State University
Columbus, Ohio
schuler@ling.osu.edu

Abstract

This paper presents a supervised pronoun anaphora resolution system based on factorial hidden Markov models (FHMMs). The basic idea is that the hidden states of FHMMs are an explicit short-term memory with an antecedent buffer containing recently described referents. Thus an observed pronoun can find its antecedent from the hidden buffer, or in terms of a generative model, the entries in the hidden buffer generate the corresponding pronouns. A system implementing this model is evaluated on the ACE corpus with promising performance.

1 Introduction

Pronoun anaphora resolution is the task of finding the correct antecedent for a given pronominal anaphor in a document. It is a subtask of coreference resolution, which is the process of determining whether two or more linguistic expressions in a document refer to the same entity. Adopting terminology used in the Automatic Context Extraction (ACE) program (NIST, 2003), these expressions are called mentions. Each mention is a reference to some entity in the domain of discourse. Mentions usually fall into three categories – proper mentions (proper names), nominal mentions (descriptions), and pronominal mentions (pronouns). There is a great deal of related work on this subject, so the descriptions of other systems below are those which are most related or which the current model has drawn insight from.

Pairwise models (Yang et al., 2004; Qiu et al., 2004) and graph-partitioning methods (McCallum and Wellner, 2003) decompose the task into a collection of pairwise or mention set coreference decisions. Decisions for each pair or each group of mentions are based on probabilities of features extracted by discriminative learning models. The aforementioned approaches have proven to be fruitful; however, there are some notable problems. Pairwise modeling may fail to produce coherent partitions. That is, if we link results of pairwise decisions to each other, there may be conflicting coreferences. Graph-partitioning methods attempt to reconcile pairwise scores into a final coherent clustering, but they are combinatorially harder to work with in discriminative approaches.

One line of research aiming at overcoming the limitation of pairwise models is to learn a mention-ranking model to rank preceding mentions for a given anaphor (Denis and Baldridge, 2007) This approach results in more coherent coreference chains.

Recent years have also seen the revival of interest in generative models in both machine learning and natural language processing. Haghighi and Klein (2007), proposed an unsupervised nonparametric Bayesian model for coreference resolution. In contrast to pairwise models, this fully generative model produces each mention from a combination of global entity properties and local attentional state. Ng (2008) did similar work using the same unsupervised generative model, but relaxed head generation as head-index generation, enforced agreement constraints at the global level, and assigned salience only to pronouns.

Another unsupervised generative model was recently presented to tackle only pronoun anaphora

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1169–1178,
Portland, Oregon, June 19-24, 2011.

resolution (Charniak and Elsner, 2009). The expectation-maximization algorithm (EM) was applied to learn parameters automatically from the parsed version of the North American News Corpus (McClosky et al., 2008). This model generates a pronoun's person, number and gender features along with the governor of the pronoun and the syntactic relation between the pronoun and the governor. This inference process allows the system to keep track of multiple hypotheses through time, including multiple different possible histories of the discourse.

Haghighi and Klein (2010) improved their non-parametric model by sharing lexical statistics at the level of abstract entity types. Consequently, their model substantially reduces semantic compatibility errors. They report the best results to date on the complete end-to-end coreference task. Further, this model functions in an online setting at mention level. Namely, the system identifies mentions from a parse tree and resolves resolution with a left-to-right sequential beam search. This is similar to Luo (2005) where a Bell tree is used to score and store the searching path.

In this paper, we present a supervised pronoun resolution system based on Factorial Hidden Markov Models (FHMMs). This system is motivated by human processing concerns, by operating incrementally and maintaining a limited short term memory for holding recently mentioned referents. According to Clark and Sengul (1979), anaphoric definite NPs are much faster retrieved if the antecedent of a pronoun is in immediately previous sentence. Therefore, a limited short term memory should be good enough for resolving the majority of pronouns. In order to construct an operable model, we also measured the average distance between pronouns and their antecedents as discussed in next sections and used distances as important salience features in the model.

Second, like Morton (2000), the current system essentially uses prior information as a discourse model with a time-series manner, using a dynamic programming inference algorithm. Third, the FHMM described here is an integrated system, in contrast with (Haghighi and Klein, 2010). The model generates part of speech tags as simple structural information, as well as related semantic information at each time step or word-by-word step. While the framework described here can be extended to deeper structural information, POS tags alone are valuable as they can be used to incorporate the binding features (described below).

Although the system described here is evaluated for pronoun resolution, the framework we describe can be extended to more general coreference resolution in a fairly straightforward manner. Further, as in other HMM-based systems, the system can be either supervised or unsupervised. But extensions to unsupervised learning are left for future work.

The final results are compared with a few supervised systems as the mention-ranking model (Denis and Baldridge, 2007) and systems compared in their paper, and Charniak and Elsner's (2009) unsupervised system, emPronouns. The FHMM-based pronoun resolution system does a better job than the global ranking technique and other approaches. This is a promising start for this novel FHMM-based pronoun resolution system.

2 Model Description

This work is based on a graphical model framework called Factorial Hidden Markov Models (FHMMs). Unlike the more commonly known Hidden Markov Model (HMM), in an FHMM the hidden state at each time step is expanded to contain more than one random variable (as shown in Figure 1). This allows for the use of more complex hidden states by taking advantage of conditional independence between substates. This conditional independence allows complex hidden states to be learned with limited training data.

2.1 Factorial Hidden Markov Model

Factorial Hidden Markov Models are an extension of HMMs (Ghahramani and Jordan, 1997). HMMs represent sequential data as a sequence of hidden states generating observation states (words in this case) at corresponding time steps t. A most likely sequence of hidden states can then be hypothesized given any sequence of observed states, using Bayes Law (Equation 2) and Markov independence assumptions (Equation 3) to define a full probability as the product of a Transition Model (Θ_T) prior probability and an Observation Model (Θ_O) likelihood probability.

$$\hat{h}_{1..T} \stackrel{\text{def}}{=} \underset{h_{1..T}}{\text{argmax}} \; \mathsf{P}(h_{1..T} \,|\, o_{1..T}) \tag{1}$$

$$\stackrel{\text{def}}{=} \underset{h_{1..T}}{\text{argmax}} \; \mathsf{P}(h_{1..T}) \cdot \mathsf{P}(o_{1..T} \,|\, h_{1..T}) \tag{2}$$

$$\stackrel{\text{def}}{=} \underset{h_{1..T}}{\text{argmax}} \prod_{t=1}^{T} \mathsf{P}_{\Theta_T}(h_t \,|\, h_{t\text{-}1}) \cdot \mathsf{P}_{\Theta_O}(o_t \,|\, h_t) \tag{3}$$

For a simple HMM, the hidden state corresponding to each observation state only involves one variable. An FHMM contains more than one hidden variable in the hidden state. These hidden substates are usually layered processes that jointly generate the evidence. In the model described here, the substates are also coupled to allow interaction between the separate processes. As Figure 1 shows, the hidden states include three sub-states, *op*, *cr* and *pos* which are short forms of *operation*, *coreference feature* and *part-of-speech*. Then, the transition model expands the left term in (3) to (4).

$$\begin{aligned} \mathsf{P}_{\Theta_T}(h_t \,|\, h_{t\text{-}1}) \stackrel{\text{def}}{=} \; & \mathsf{P}(op_t \,|\, op_{t-1}, pos_{t-1}) \\ & \cdot \mathsf{P}(cr_t \,|\, cr_{t-1}, op_{t-1}) \\ & \cdot \mathsf{P}(pos_t \,|\, op_t, pos_{t-1}) \end{aligned} \tag{4}$$

The observation model expands from the right term in (3) to (5).

$$\mathsf{P}_{\Theta_O}(o_t \,|\, h_t) \stackrel{\text{def}}{=} \mathsf{P}(o_t \,|\, pos_t, cr_t) \tag{5}$$

The observation state depends on more than one hidden state at each time step in an FHMM. Each hidden variable can be further split into smaller variables. What these terms stand for and the motivations behind the above equations will be explained in the next section.

2.2 Modeling a Coreference Resolver with FHMMs

FHMMs in our model, like standard HMMs, cannot represent the hierarchical structure of a syntactic phrase. In order to partially represent this information, the head word is used to represent the whole noun phrase. After coreference is resolved, the coreferring chain can then be expanded to the whole phrase with NP chunker tools.

In this system, hidden states are composed of three main variables: a referent operation (*OP*), coreference features (*CR*) and part of speech tags (*POS*) as displayed in Figure 1. The transition model is defined as Equation 4.

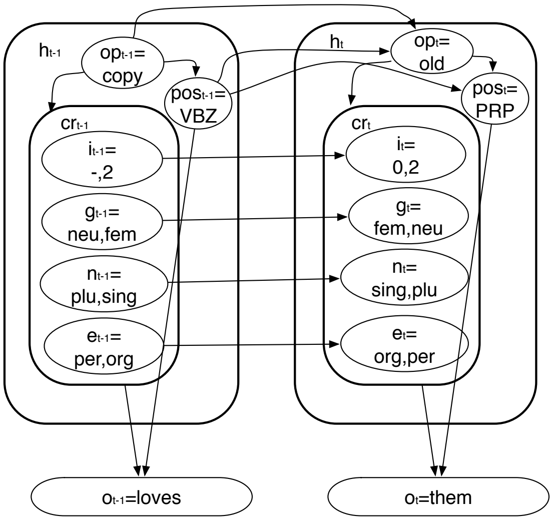

Figure 1: Factorial HMM CR Model

The starting point for the hidden state at each time step is the *OP* variable, which determines which kind of referent operations will occur at the current word. Its domain has three possible states: *none*, *new* and *old*.

The *none* state indicates that the present state will not generate a mention. All previous hidden state values (the list of previous mentions) will be passed deterministically (with probability 1) to the current time step without any changes. The *new* state signifies that there is a new mention in the present time step. In this event, a new mention will be added to the entity set, as represented by its set of feature values and position in the coreference table. The *old* state indicates that there is a mention in the present time state and that this mention refers back to some antecedent mention. In such a case, the list of entities in the buffer will be reordered deterministically, moving the currently mentioned entity to the top of the list.

Notice that op_t is defined to depend on op_{t-1} and pos_{t-1}. This is sometimes called a *switching* FHMM (Duh, 2005). This dependency can be useful, for example, if op_{t-1} is *new*, in which case op_t has a higher probability of being *none* or *old*. If

pos_{t-1} is a verb or preposition, op_t has more probability of being *old* or *new*.

One may wonder why op_t generates pos_t, and not the other way around. This model only roughly models the process of (new and old) entity generation, and either direction of causality might be consistent with a model of human entity generation, but this direction of causality is chosen to represent the effect of semantics (referents) generating syntax (POS tags). In addition, this is a joint model in which POS tagging and coreference resolution are integrated together, so the best combination of those hidden states will be computed in either case.

2.3 Coreference Features

Coreference features for this model refer to features that may help to identify co-referring entities.

In this paper, they mainly include index (*I*), named entity type (*E*), number (*N*) and gender (*G*). The index feature represents the order that a mention was encountered relative to the other mentions in the buffer. The latter three features are well known and described elsewhere, and are not themselves intended as the contribution of this work. The novel aspect of this part of the model is the fact that the features are carried forward, updated after every word, and essentially act as a discourse model. The features are just a shorthand way of representing some well known essential aspects of a referent (as pertains to anaphora resolution) in a discourse model.

Features	Values
I	positive integers from 1. . .n
G	male, female, neutral, unknown
N	singular, plural, unknown
E	person, location, organization, GPE, vehicle, company, facility

Table 1: Coreference features stored with each mention.

Unlike discriminative approaches, generative models like the FHMM described here do not have access to all observations at once. This model must then have a mechanism for jointly considering pronouns in tandem with previous mentions, as well as the features of those mentions that might be used to find matches between pronouns and antecedents.

Further, higher order HMMs may contain more accurate information about observation states. This is especially true for coreference resolution because pronouns often refer back to mentions that are far away from the present state. In this case, we would need to know information about mentions which are at least two mentions before the present one. In this sense, a higher order HMM may seem ideal for coreference resolution. However, higher order HMMs will quickly become intractable as the order increases.

In order to overcome these limitations, two strategies which have been discussed in the last section are taken: First, a switching variable called *OP* is designed (as discussed in last section); second, a memory of recently mentioned entities is maintained to store features of mentions and pass them forward incrementally.

OP is intended to model the decision to use the current word to introduce a new referent (*new*), refer to an antecedent (*old*), or neither (*none*). The entity buffer is intended to model the set of ‘activated’ entities in the discourse – those which could plausibly be referred to with a pronoun. These designs allow similar benefits as longer dependencies of higher-order HMMs but avoid the problem of intractability. The number of mentions maintained must be limited in order for the model to be tractable. Fortunately, human short term memory faces effectively similar limitations and thus pronouns usually refer back to mentions not very far away.

Even so, the impact of the size of the buffer on decoding time may be a concern. Since the buffer of our system will carry forward a few previous groups of coreference features plus *op* and *pos*, the computational complexity will be exorbitantly high if we keep high beam size and meanwhile if each feature interacts with others. Luckily, we have successfully reduced the intractability to a workable system in both speed and space with following methods. First, we estimate the size of buffer with a simple count of average distances between pronouns and their antecedents in the corpus. It is found that about six is enough for covering 99.2% of all pronouns.

Secondly, the coreference features we have used have the nice property of being independent from one another. One might expect English non-person entities to almost always have neutral gender, and thus be modeled as follows:

$$\mathsf{P}(e_t, g_t \,|\, e_{t-1}, g_{t-1}) = \mathsf{P}(g_t \,|\, g_{t-1}, e_t) \cdot \mathsf{P}(e_t \,|\, e_{t-1}) \tag{6}$$

However, a few considerations made us reconsider. First, exceptions are found in the corpus. Personal pronouns such as *she* or *he* are used to refer to country, regions, states or organizations. Second, existing model files made by Bergsma (2005) include a large number of non-neutral gender information for non-person words. We employ these files for acquiring gender information of unknown words. If we use Equation 6, sparsity and complexity will increase. Further, preliminary experiments have shown models using an independence assumption between gender and personhood work better. Thus, we treat each coreference feature as an independent event. Hence, we can safely split coreference features into separate parts. This way dramatically reduces the model complexity. Thirdly, our HMM decoding uses the Viterbi algorithm with A-star beam search.

The probability of the new state of the coreference table $\mathsf{P}(cr_t \,|\, cr_{t-1}, op_t)$ is defined to be the product of probabilities of the individual feature transitions.

$$\begin{aligned} \mathsf{P}(cr_t \,|\, cr_{t-1}, op_t) = \mathsf{P}(i_t \,|\, i_{t-1}, op_t)\cdot \\ \mathsf{P}(e_t \,|\, e_{t-1}, i_t, op_t)\cdot \\ \mathsf{P}(g_t \,|\, g_{t-1}, i_t, op_t)\cdot \\ \mathsf{P}(n_t \,|\, n_{t-1}, i_t, op_t) \end{aligned} \tag{7}$$

This supposes that the features are conditionally independent of each other given the index variable, the operator and previous instance. Each feature only depends on the operator and the corresponding feature at the previous state, with that set of features re-ordered as specified by the index model.

2.4 Feature Passing

Equation 7 is correct and complete, but in fact the switching variable for operation type results in three different cases which simplifies the calculation of the transition probabilities for the coreference feature table.

Note the following observations about coreference features: i_t only needs a probabilistic model when op_t is *old* – in other words, only when the model must choose between several antecedents to re-refer to. g_t, e_t and n_t are deterministic except when op_t is *new*, when gender, entity type, and number information must be generated for the new entity being introduced.

When op_t is *none*, all coreference variables (entity features) will be copied over from the previous time step to the current time step, and the probability of this transition is 1.0. When op_t is *new*, i_t is changed deterministically by adding the new entity to the first position in the list and moving every other entity down one position. If the list of entities is full, the least recently mentioned entity will be discarded. The values for the top of the feature lists g_t, e_t, and n_t will then be generated from feature-specific probability distributions estimated from the training data. When op_t is *old*, i_t will probabilistically select a value $1 \ldots n$, for an entity list containing n items. The selected value will deterministically order the g_t, n_t and e_t lists. This distribution is also estimated from training data, and takes into account recency of mention. The shape of this distribution varies slightly depending on list size and noise in the training data, but in general the probability of a mention being selected is directly correlated to how recently it was mentioned.

With this understanding, coreference table transition probabilities can be written in terms of only their non-deterministic substate distributions:

$$\begin{aligned} \mathsf{P}(cr_t \,|\, cr_{t-1}, old) = \; & \mathsf{P}_{old}(i_t \,|\, i_{t-1})\cdot \\ & \mathsf{P}_{reorder}(e_t \,|\, e_{t-1}, i_t)\cdot \\ & \mathsf{P}_{reorder}(g_t \,|\, g_{t-1}, i_t)\cdot \\ & \mathsf{P}_{reorder}(n_t \,|\, n_{t-1}, i_t) \end{aligned} \tag{8}$$

where the *old* model probabilistically selects the antecedent and moves it to the top of the list as described above, thus deciding how the reordering will take place. The *reorder* model actually implements the list reordering for each independent feature by moving the feature value corresponding to the selected entity in the index model to the top of that feature's list. The overall effect is simply the probabilistic reordering of entities in a list, where each entity is defined as a label and a set of features.

$$\begin{aligned} \mathsf{P}(cr_t \,|\, cr_{t-1}, new) = \; & \mathsf{P}_{new}(i_t \,|\, i_{t-1})\cdot \\ & \mathsf{P}_{new}(g_t \,|\, g_{t-1})\cdot \\ & \mathsf{P}_{new}(n_t \,|\, n_{t-1})\cdot \\ & \mathsf{P}_{new}(e_t \,|\, e_{t-1}) \end{aligned} \tag{9}$$

where the *new* model probabilistically generates a feature value based on the training data and puts it at the top of the list, moves every other entity down one position in the list, and removes the final item if the list is already full. Each entity in *i* takes a value from 1 to n for a list of size n. Each *g* can be one of four values – *male*, *female*, *neuter* and *unknown*; *n* one of three values – *plural*, *singular* and *unknown* and *e* around eight values.

Note that pos_t is used in both hidden states and observation states. While it is not considered a coreference feature as such, it can still play an important role in the resolving process. Basically, the system tags parts of speech incrementally while simultaneously resolving pronoun anaphora. Meanwhile, pos_{t-1} and op_{t-1} will jointly generate op_t. This point has been discussed in Section 2.2.

Importantly, the pos model can help to implement binding principles (Chomsky, 1981). It is applied when op_t is *old*. In training, pronouns are sub-categorised into personal pronouns, reflexive and other-pronoun. We then define a variable loc_t whose value is how far back in the list of antecedents the current hypothesis must have gone to arrive at the current value of i_t. If we have the syntax annotations or parsed trees, then, the part of speech model can be defined when op_t is *old* as $\mathsf{P}_{binding}(pos_t \,|\, loc_t, s_{loc_t})$. For example, if $pos_t \in reflexive$, $\mathsf{P}(pos_t \,|\, loc_t, s_{loc_t})$ where loc_t has smaller values (implying closer mentions to pos_t) and $s_{loc_t} = subject$ should have higher values since reflexive pronouns always refer back to subjects within its governing domains. This was what (Haghighi and Klein, 2009) did and we did this in training with the REUTERS corpus (Hasler et al., 2006) in which syntactic roles are annotated. We finally switched to the ACE corpus for the purpose of comparison with other work. In the ACE corpus, no syntactic roles are annotated. We did use the Stanford parser to extract syntactic roles from the ACE corpus. But the result is largely affected by the parsing accuracy. Again, for a fair comparison, we extract similar features to Denis and Baldridge (2007), which is the model we mainly compare with. They approximate syntactic contexts with POS tags surrounding the pronoun. Inspired by this idea, we successfully represent binding features with POS tags before anaphors. Instead of using $\mathsf{P}(pos_t \,|\, loc_t, s_{loc_t})$, we train $\mathsf{P}(pos_t \,|\, loc_t, pos_{loc_t})$ which can play the role of binding. For example, suppose the buffer size is 6 and $loc_t = 5$, $pos_{loc_t} = noun$. Then, $\mathsf{P}(pos_t = reflexive \,|\, loc_t, pos_{loc_t})$ is usually higher than $\mathsf{P}(pos_t = pronoun \,|\, loc_t, pos_{loc_t})$, since the reflexive has a higher probability of referring back to the noun located in position 5 than the pronoun.

In future work expanding to coreference resolution between any noun phrases we intend to integrate syntax into this framework as a joint model of coreference resolution and parsing.

3 Observation Model

The observation model that generates an observed state is defined as Equation 5. To expand that equation in detail, the observation state, the word, depends on its part of speech and its coreference features as well. Since FHMMs are generative, we can say part of speech and coreference features generate the word.

In actual implementation, the observed model will be very sparse, since cr_t will be split into more variables according to how many coreference features it is composed of. In order to avoid the sparsity, we transform the equation with Bayes' law as follows.

$$\mathsf{P}_{\Theta_O}(o_t \,|\, h_t) = \frac{P(o_t) \cdot \mathsf{P}(h_t \,|\, o_t)}{\sum_{o'} P(o')\mathsf{P}(h_t \,|\, o')} \tag{10}$$

$$= \frac{P(o_t) \cdot \mathsf{P}(pos_t, cr_t \,|\, o_t)}{\sum_{o'} P(o')\mathsf{P}(pos_t, cr_t \,|\, o')} \tag{11}$$

We define *pos* and *cr* to be independent of each other, so we can further split the above equation as:

$$\mathsf{P}_{\Theta_O}(o_t \,|\, h_t) \overset{\text{def}}{=} \frac{P(o_t) \cdot \mathsf{P}(pos_t \,|\, o_t) \cdot \mathsf{P}(cr_t \,|\, o_t)}{\sum_{o'} P(o') \cdot \mathsf{P}(pos_t \,|\, o') \cdot \mathsf{P}(cr_t \,|\, o')} \tag{12}$$

where $\mathsf{P}(cr_t \,|\, o_t) = \mathsf{P}(g_t \,|\, o_t)\mathsf{P}(n_t \,|\, o_t)\mathsf{P}(e_t \,|\, o_t)$ and $\mathsf{P}(cr_t \,|\, o') = \mathsf{P}(g_t \,|\, o')\mathsf{P}(n_t \,|\, o')\mathsf{P}(e_t \,|\, o')$.

This change transforms the FHMM to a hybrid FHMM since the observation model no longer generates the data. Instead, the observation model generates hidden states, which is more a combination of discriminative and generative approaches. This way facilitates building likelihood model files of features for given mentions from the training data. The hidden state transition model represents prior probabilities of coreference features associated with each while this observation model factors in the probability given a pronoun.

3.1 Unknown Words Processing

If an observed word was not seen in training, the distribution of its part of speech, gender, number and entity type will be unknown. In this case, a special unknown words model is used.

The part of speech of unknown words $\mathsf{P}(pos_t \,|\, w_t = unkword)$ is estimated using a decision tree model. This decision tree is built by splitting letters in words from the end of the word backward to its beginning. A POS tag is assigned to the word after comparisons between the morphological features of words trained from the corpus and the strings concatenated from the tree leaves are made. This method is about as accurate as the approach described by Klein and Manning (2003).

Next, a similar model is set up for estimating $\mathsf{P}(n_t \,|\, w_t = unkword)$. Most English words have regular plural forms, and even irregular words have their patterns. Therefore, the morphological features of English words can often be used to determine whether a word is singular or plural.

Gender is irregular in English, so model-based predictions are problematic. Instead, we follow Bergsma and Lin (2005) to get the distribution of gender from their gender/number data and then predict the gender for unknown words.

4 Evaluation and Discussion

4.1 Experimental Setup

In this research, we used the ACE corpus (Phase 2) [1] for evaluation. The development of this corpus involved two stages. The first stage is called EDT (entity detection and tracking) while the second stage is called RDC (relation detection and characterization). All markables have named entity types such as FACILITY, GPE (geopolitical entity), PERSON, LOCATION, ORGANIZATION, PERSON, VEHICLE and WEAPONS, which were annotated in the first stage. In the second stage, relations between named entities were annotated. This corpus include three parts, composed of different genres: newspaper texts (NPAPER), newswire texts (NWIRE) and broadcasted news (BNEWS). Each of these is split into a *train* part and a *devtest* part. For the train part, there are 76, 130 and 217 articles in NPAPER, NWIRE and BNEWS respectively while for the test part, there are 17, 29 and 51 articles respectively. Though the number of articles are quite different for three genres, the total number of words are almost the same. Namely, the length of NPAPER is much longer than BNEWS (about 1200 words, 800 word and 500 words respectively for three genres). The longer articles involve longer coreference chains. Following the common practice, we used the *devtest* material only for testing. Progress during the development phase was estimated only by using cross-validation on the training set for the BNEWS section. In order to make comparisons with publications which used the same corpus, we make efforts to set up identical conditions for our experiments.

The main point of comparison is Denis and Baldridge (2007), which was similar in that it described a new type of coreference resolver using simple features.

Therefore, similar to their practice, we use all forms of personal and possessive pronouns that were annotated as ACE ”markables”. Namely, pronouns associated with named entity types could be used in this system. In experiments, we also used *true* ACE mentions as they did. This means that pleonastics and references to eventualities or to non-ACE entities are not included in our experiments either. In all, 7263 referential pronouns in training data set and 1866 in testing data set are found in all three genres. They have results of three different systems: SCC (single candidate classifier), TCC (twin candidate classifier) and RK (ranking). Besides the three and our own system, we also report results of emPronouns, which is an unsupervised system based on a recently published paper (Charniak and Elsner, 2009). We select this unsupervised system for two reasons. Firstly, emPronouns is a publicly available system with high accuracy in pronoun resolution. Secondly, it is necessary for us to demonstrate our system has strong empirical superiority over unsupervised ones. In testing, we also used the OPNLP Named Entity Recognizer to tag the test corpus.

[1] See `http://projects.ldc.upenn.edu/ace/annotation/previous/` for details on the corpus.

During training, besides coreference annotation itself, the part of speech, dependencies between words and named entities, gender, number and index are extracted using relative frequency estimation to train models for the coreference resolution system.

Inputs for testing are the plain text and the trained model files. The entity buffer used in these experiments kept track of only the six most recent mentions. The result of this process is an annotation of the headword of every noun phrase denoting it as a mention. In addition, this system does not do anaphoricity detection, so the antecedent operation for non-anaphora pronoun *it* is set to be *none*. Finally, the system does not yet model cataphora, about 10 cataphoric pronouns in the testing data which are all counted as wrong.

4.2 Results

The performance was evaluated using the ratio of the number of correctly resolved anaphors over the number of all anaphors as a success metrics. All the standards are consistent with those defined in Charniak and Elsner (2009).

During development, several preliminary experiments explored the effects of starting from a simple baseline and adding more features. The BNEWS corpus was employed in these development experiments. The baseline only includes part of speech tags, the index feature and and syntactic roles. Syntactic roles are extracted from the parsing results with Stanford parser. The success rate of this baseline configuration is 0.48. This low accuracy is partially due to the errors of automatic parsing. With gender and number features added, the performance jumped to 0.65. This shows that number and gender agreements play an important role in pronoun anaphora resolution. For a more standard comparison to other work, subsequent tests were performed on the gold standard ACE corpus (using the model as described with named entity features instead of syntactic role features). As shown in Denis and Baldridge (2007), they employ all features we use except syntactic roles. In these experiments, the system got better results as shown in Table 2.

System	BNEWS	NPAPER	NWIRE
emPronouns	58.5	64.5	60.6
SCC	62.2	70.7	68.3
TCC	68.6	74.7	71.1
RK	72.9	76.4	72.4
FHMM	**74.9**	**79.4**	**74.5**

Table 2: Accuracy scores for emPronouns, the single-candidate classifier (SCC), the twin-candidate classifier (TCC), the ranker and FHMM

The result of the first one is obtained by running the publicly available system emPronouns[2]. It is a high-accuracy unsupervised system which reported the best result in Charniak and Elsner (2009).

The results of the other three systems are those reported by Denis and Baldridge (2007). As Table 2 shows, the FHMM system gets the highest average results.

The emPronouns system got the lowest results partially due to the reason that we only directly run the existing system with its existing model files without retraining. But the gap between its results and results of our system is large. Thus, we may still say that our system probably can do a better job even if we train new models files for emPronouns with ACE corpus.

With almost exactly identical settings, why does our FHMM system get the highest average results? The convincing reason is that FHMM is strongly influenced by the sequential dependencies. The ranking approach ranks a set of mentions using a set of features, and it also maintains the discourse model, but it is not processing sequentially. The FHMM system always maintain a set of mentions as well as a first-order dependencies between part of speech and operator. Therefore, context can be more fully taken into consideration. This is the main reason that the FHMM approach achieved better results than the ranking approach.

From the result, one point we may notice is that NPAPER usually obtains higher results than both BNEWS and NWIRE for all systems while BNEWS lower than other two genres. In last section, we mention that articles in NPAPER are longer than other genres and also have denser coreference chains while articles in BENEWS are shorter and have sparer chains. Then, it is not hard to understand why results of NPAPER are better while those of

[2] the available system in fact only includes the testing part. Thus, it may be unfair to compare emPronouns this way with other systems.

BNEWS are poorer.

In Denis and Baldridge (2007), they also reported new results with a window of 10 sentences for RK model. All three genres obtained higher results than those when with shorter ones. They are 73.0, 77.6 and 75.0 for BNEWS, NPAPER and NWIRE respectively. We can see that except the one for NWIRE, the results are still poorer than our system. For NWIRE, the RK model got 0.5 higher. The average of the RK is 75.2 while that of the FHMM system is 76.3, which is still the best.

Since the emPronoun system can output sample-level results, it is possible to do a paired Student's t-test. That test shows that the improvement of our system on all three genres is statistically significant ($p < 0.001$). Unfortunately, the other systems only report overall results so the same comparison was not so straightforward.

4.3 Error Analysis

After running the system on these documents, we checked which pronouns fail to catch their antecedents. There are a few general reasons for errors.

First, pronouns which have antecedents very far away cannot be caught. Long-distance anaphora resolution may pose a problem since the buffer size cannot be too long considering the complexity of tracking a large number of mentions through time. During development, estimation of an acceptable size was attempted using the training data. It was found that a mention distance of fourteen would account for every case found in this corpus, though most cases fall well short of that distance. Future work will explore optimizations that will allow for larger or variable buffer sizes so that longer distance anaphora can be detected.

A second source of error is simple misjudgments when more than one candidate is waiting for selection. A simple case is that the system fails to distinguish plural personal nouns and non-personal nouns if both candidates are plural. This is not a problem for singular pronouns since gender features can tell whether pronouns are personal or not. Plural nouns in English do not have such distinctions, however. Consequently, *demands* and *Israelis* have the same probability of being selected as the antecedents for *they*, all else being equal. If *demands* is closer to *they*, *demands* will be selected as the antecedent. This may lead to the wrong choice if *they* in fact refers to *Israelis*. This may require better measures of referent salience than the "least recently used" heuristic currently implemented.

Third, these results also show difficulty resolving coordinate noun phrases due to the simplistic representation of noun phrases in the input. Consider this sentence: *President Barack Obama and his wife Michelle Obama visited China last week. They had a meeting with President Hu in Beijing*. In this example, the pronoun *they* corefers with the noun phrase *President Barack Obama and his wife Michelle Obama*. The present model cannot represent both the larger noun phrase and its contained noun phrases. Since the noun phrase is a coordinate one that includes both noun phrases, the model cannot find a head word to represent it.

Finally, while the coreference feature annotations of the ACE are valuable for learning feature models, the model training may still give some misleading results. This is brought about by missing features in the training corpus and by the data sparsity. We solved the problem with add-one smoothing and deleted interpolation in training models besides the transformation in the generation order of the observation model.

5 Conclusion and Future Work

This paper has presented a pronoun anaphora resolution system based on FHMMs. This generative system incrementally resolves pronoun anaphora with an entity buffer carrying forward mention features. The system performs well and outperforms other available models. This shows that FHMMs and other time-series models may be a valuable model to resolve anaphora.

Acknowledgments

We would like to thank the authors and maintainers of ranker models and emPronouns. We also would like to thank the three anonymous reviewers. The final version is revised based on their valuable comments. Thanks are extended to Shane Bergsma, who provided us the gender and number data distribution. In addition, Professor Jeanette Gundel and our labmate Stephen Wu also gave us support in paper editing and in theoretical discussion.

References

S Bergsma. 2005. Automatic acquisition of gender information for anaphora resolution. page 342353. Springer.

Eugene Charniak and Micha Elsner. 2009. Em works for pronoun anaphora resolution. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, Athens, Greece.

Noam Chomsky. 1981. *Lectures on government and binding*. Foris, Dordercht.

H.H. Clark and CJ Sengul. 1979. In search of referents for nouns and pronouns. *Memory & Cognition*, 7(1):35–41.

P. Denis and J. Baldridge. 2007. A ranking approach to pronoun resolution. In *Proc. IJCAI*.

Kevin Duh. 2005. Jointly labeling multiple sequences: a factorial HMM approach. In *ACL '05: Proceedings of the ACL Student Research Workshop*, pages 19–24, Ann Arbor, Michigan.

Zoubin Ghahramani and Michael I. Jordan. 1997. Factorial hidden markov models. *Machine Learning*, 29:1–31.

A. Haghighi and D. Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th annual meeting on Association for Computational Linguistics*, page 848.

A. Haghighi and D. Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1152–1161. Association for Computational Linguistics.

A. Haghighi and D. Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

L. Hasler, C. Orasan, and K. Naumann. 2006. NPs for events: Experiments in coreference annotation. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006)*, pages 1167–1172. Citeseer.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.

X Luo. 2005. On coreference resolution performance metrics. pages 25–32. Association for Computational Linguistics Morristown, NJ, USA.

A. McCallum and B. Wellner. 2003. Toward conditional models of identity uncertainty with application to proper noun coreference. In *IJCAI Workshop on Information Integration on the Web*. Citeseer.

David McClosky, Eugene Charniak, and Mark Johnson. 2008. BLLIP North American News Text, Complete. *Linguistic Data Consortium. LDC2008T13*.

T.S. Morton. 2000. Coreference for NLP applications. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics.

V. Ng. 2008. Unsupervised models for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 640–649. Association for Computational Linguistics.

US NIST. 2003. The ACE 2003 Evaluation Plan. *US National Institute for Standards and Technology (NIST), Gaithersburg, MD.[online*, pages 2003–08.

L. Qiu, M.Y. Kan, and T.S. Chua. 2004. A public reference implementation of the rap anaphora resolution algorithm. *Arxiv preprint cs/0406031*.

X. Yang, J. Su, G. Zhou, and C.L. Tan. 2004. Improving pronoun resolution by incorporating coreferential information of candidates. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 127. Association for Computational Linguistics.