Structural, Transitive and Latent Models for Biographic Fact Extraction

Nikesh Garera and **David Yarowsky**
Department of Computer Science, Johns Hopkins University
Human Language Technology Center of Excellence
Baltimore MD, USA
{ngarera,yarowsky}@cs.jhu.edu

Abstract

This paper presents six novel approaches to biographic fact extraction that model structural, transitive and latent properties of biographical data. The ensemble of these proposed models substantially outperforms standard pattern-based biographic fact extraction methods and performance is further improved by modeling inter-attribute correlations and distributions over functions of attributes, achieving an average extraction accuracy of 80% over seven types of biographic attributes.

1 Introduction

Extracting biographic facts such as "Birthdate", "Occupation", "Nationality", etc. is a critical step for advancing the state of the art in information processing and retrieval. An important aspect of web search is to be able to narrow down search results by distinguishing among people with the same name leading to multiple efforts focusing on web person name disambiguation in the literature (Mann and Yarowsky, 2003; Artiles et al., 2007, Cucerzan, 2007). While biographic facts are certainly useful for disambiguating person names, they also allow for automatic extraction of encyclopedic knowledge that has been limited to manual efforts such as Britannica, Wikipedia, etc. Such encyploedic knowledge can advance vertical search engines such as http://www.spock.com that are focused on people searches where one can get an enhanced search interface for searching by various biographic attributes. Biographic facts are also useful for powerful query mechanisms such as finding what attributes are common between two people (Auer and Lehmann, 2007).

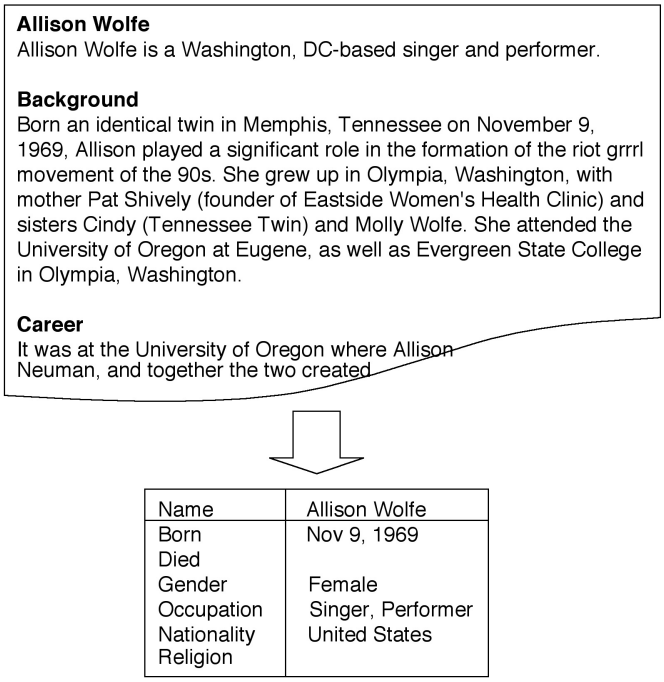

Figure 1: Goal: extracting attribute-value biographic fact pairs from biographic free-text

While there are a large quantity of biographic texts available online, there are only a few biographic fact databases available[1], and most of them have been created manually, are incomplete and are available primarily in English.

This work presents multiple novel approaches for automatically extracting biographic facts such as "Birthdate", "Occupation", "Nationality", and "Religion", making use of diverse sources of information present in biographies.

In particular, we have proposed and evaluated the following 6 distinct original approaches to this

[1] E.g.: http://www.nndb.com, http://www.biography.com, Infoboxes in Wikipedia

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 300–308,
Athens, Greece, 30 March – 3 April 2009.

task with large collective empirical gains:

1. An improvement to the Ravichandran and Hovy (2002) algorithm based on *Partially Untethered Contextual Pattern Models*

2. Learning a *position-based* model using absolute and relative positions and sequential order of hypotheses that satisfy the domain model. For example, "Deathdate" very often appears after "Birthdate" in a biography.

3. Using *transitive models over attributes* via co-occurring entities. For example, other people mentioned person's biography page tend to have similar attributes such as occupation (See Figure 4).

4. Using *latent wide-document-context models* to detect attributes that may not be mentioned directly in the article (e.g. the words *"song, hits, album, recorded,.."* all collectively indicate the occupation of *singer* or *musician* in the article.

5. Using *inter-attribute correlations*, for filtering unlikely biographic attribute combinations. For example, a tuple consisting of < "Nationality" = India, "Religion" = Hindu > has a higher probability than a tuple consisting of < "Nationality" = France, "Religion" = Hindu >.

6. Learning *distributions over functions of attributes*, for example, using an age distribution to filter tuples containing improbable <deathyear>-<birthyear> lifespan values.

We propose and evaluate techniques for exploiting all of the above classes of information in the next sections.

2 Related Work

The literature for biography extraction falls into two major classes. The first one deals with identifying and extracting *biographical sentences* and treats the problem as a summarization task (Cowie et al., 2000, Schiffman et al., 2001, Zhou et al., 2004). The second and more closely related class deals with extracting specific *facts* such as *"birthplace", "occupation", etc*. For this task, the primary theme of work in the literature has been to treat the task as a general semantic-class learning problem where one starts with a few seeds of the semantic relationship of interest and learns contextual patterns such as "<NAME> was born in <Birthplace>" or "<NAME> (born <Birthdate>)" (Hearst, 1992; Riloff, 1996; Thelen and Riloff, 2002; Agichtein and Gravano, 2000; Ravichandran and Hovy, 2002; Mann and Yarowsky, 2003; Jijkoun et al., 2004; Mann and Yarowsky, 2005; Alfonseca et al., 2006; Pasca et al., 2006). There has also been some work on extracting biographic facts directly from Wikipedia pages. Culotta et al. (2006) deal with learning contextual patterns for extracting family relationships from Wikipedia. Ruiz-Casado et al. (2006) learn contextual patterns for biographic facts and apply them to Wikipedia pages.

While the pattern-learning approach extends well for a few biography classes, some of the biographic facts like "Gender" and "Religion" do not have consistent contextual patterns, and only a few of the explicit biographic attributes such as "Birthdate", "Deathdate", "Birthplace" and "Occupation" have been shown to work well in the pattern-learning framework (Mann and Yarowsky, 2005; Alfonesca, 2006; Pasca et al., 2006).

Secondly, there is a general lack of work that attempts to utilize the typical information sequencing within biographic texts for fact extraction, and we show how the information structure of biographies can be used to improve upon pattern based models. Furthermore, we also present additional novel models of attribute correlation and age distribution that aid the extraction process.

3 Approach

We first implement the standard pattern-based approach for extracting biographic facts from the raw prose in Wikipedia people pages. We then present an array of novel techniques exploiting different classes of information including partially-tethered contextual patterns, relative attribute position and sequence, transitive attributes of co-occurring entities, broad-context topical profiles, inter-attribute correlations and likely human age distributions. For illustrative purposes, we motivate each technique using one or two attributes but in practice they can be applied to a wide range of attributes and empirical results in Table 4 show that they give consistent performance gains across multiple attributes.

4 Contextual Pattern-Based Model

A standard model for extracting biographic facts is to learn templatic contextual patterns such as <NAME> "was born in" <Birthplace>. Such templatic patterns can be learned using seed examples of the attribute in question and, there has been a plethora of work in the seed-based bootstrapping literature which addresses this problem (Ravichandran and Hovy, 2002; Thelen and Riloff, 2002; Mann and Yarowsky, 2005; Alfonseca et al., 2006; Pasca et al., 2006)

Thus for our baseline we implemented a standard Ravichandran and Hovy (2002) pattern learning model using 100 seed[2] examples from an online biographic database called NNDB (http://www.nndb.com) for each of the biographic attributes: "Birthdate", "Birthplace", "Deathdate", "Gender", "Nationality", "Occupation" and "Religion". Given the seed pairs, patterns for each attribute were learned by searching for seed *<Name,Attribute Value>* pairs in the Wikipedia page and extracting the left, middle and right contexts as various contextual patterns[3].

While the biographic text was obtained from Wikipedia articles, all of the 7 attribute values used as seed and test person names could not be obtained from Wikipedia due to incomplete and unnormalized (for attribute value format) infoboxes. Hence, the values for training/evaluation were extracted from NNDB which provides a cleaner set of gold truth, and is similar to an approach utilizing trained annotators for marking up and extracting the factual information in a standard format. For consistency, only the people names whose articles occur in Wikipedia where selected as part of seed and test sets.

Given the attribute values of the seed names and their text articles, the probability of a relationship r(Attribute Name), given the surrounding context "A_1 p A_2 q A_3", where p and q are <NAME> and <Attrib Val> respectively, is given using the rote extractor model probability as in (Ravichandran and Hovy, 2002; Mann and Yarowsky 2005):

$$P(r(p,q)|A_1pA_2qA_3) = \frac{\sum_{x,y \in r} c(A_1xA_2yA_3)}{\sum_{x,z} c(A_1xA_2zA_3)}$$

Thus, the probability for each contextual pattern is based on how often it correctly predicts a relationship in the seed set. And, each extracted attribute value q using the given pattern can thus be ranked according to the above probability. We tested this approach for extracting values for each of the seven attributes on a test set of 100 held-out names and report Precision, Pseudo-recall and F-score for each attribute which are computed in the standard way as follows, for say Attribute "Birthplace (bplace)":

$$\text{Precision}_{\text{bplace}} = \frac{\text{\# people with bplace correctly extracted}}{\text{\# of people with bplace extracted}}$$

$$\text{Pseudo-rec}_{\text{bplace}} = \frac{\text{\# people with bplace correctly extracted}}{\text{\# of people with bplace in test set}}$$

$$\text{F-score}_{\text{bplace}} = \frac{2 \cdot \text{Precision}_{\text{bplace}} \cdot \text{Pseudo-rec}_{\text{bplace}}}{\text{Precision}_{\text{bplace}} + \text{Pseudo-rec}_{\text{bplace}}}$$

Since the true values of each attribute are obtained from a cleaner and normalized person-database (NNDB), not all the attribute values maybe present in the Wikipedia article for a given name. Thus, we also compute accuracy on the subset of names for which the value of a given attribute is also explictly stated in the article. This is denoted as:

$$\text{Acc}_{\text{truth pres}} = \frac{\text{\# people with bplace correctly extracted}}{\text{\# of people with true bplace stated in article}}$$

We further applied a domain model for each attribute to filter noisy targets extracted from lexical patterns. Our domain models of attributes include lists of acceptable values (such as lists of places, occupations and religions) and structural constraints such as possible date formats for "Birthdate" and "Deathdate". The rows with subscript "$RH02$"in Table 4 shows the performance of this Ravichandran and Hovy (2002) model with additional attribute domain modeling for each attribute, and Table 3 shows the average performance across all attributes.

5 Partially Untethered Templatic Contextual Patterns

The pattern-learning literature for fact extraction often consists of patterns with a "hook" and "target" (Mann and Yarowsky, 2005). For example, in the pattern "<Name> *was born in* <Birthplace>", "<NAME>" is the hook and "<Birthplace>" is the target. The disadvantage of this approach is that the intervening dually-tethered patterns can be quite long and highly variable, such as "<NAME> *was highly influ-*

[2]The seed examples were chosen randomly, with a bias against duplicate attribute values to increase training diversity. Both the seed and test names and data will be made available online to the research community for replication and extension.

[3]We implemented a noisy model of coreference resolution by resolving any gender-correct pronoun used in the Wikipedia page to the title person name of the article. Gender is also extracted automatically as a biographic attribute.

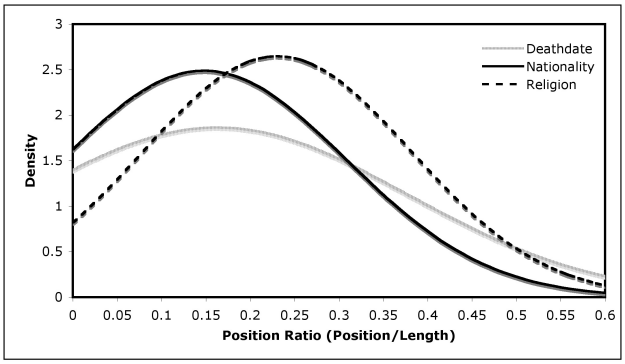

Figure 2: Distribution of the observed document mentions of Deathdate, Nationality and Religion.

ential in his role as <Occupation>". We overcome this problem by modeling partially untethered variable-length ngram patterns adjacent to only the *target*, with the only constraint being that the hook entity appear somewhere in the sentence[4]. Examples of these new contextual ngram features include "*his role as* <Occupation>" and '*role as* <Occupation>". The pattern probability model here is essentially the same as in Ravichandran and Hovy, 2002 and just the pattern representation is changed. The rows with subscript "$RH02_{\text{imp}}$" in tables 4 and 3 show performance gains using this improved templatic-pattern-based model, yielding an absolute 21% gain in accuracy.

6 Document-Position-Based Model

One of the properties of biographic genres is that primary biographic attributes[5] tend to appear in characteristic positions, often toward the beginning of the article. Thus, the absolute position (in percentage) can be modeled explicitly using a Gaussian parametric model as follows for choosing the best candidate value v^* for a given attribute A:

$$v^* = argmax_{v \in domain(A)} f(posn_v|A)$$

where,

$$f(posn_v|A)$$
$$= \mathcal{N}(posn_v; \hat{\mu_A}, \hat{\sigma_A^2})$$
$$= \frac{1}{\hat{\sigma_A}\sqrt{2\pi}} e^{-(posn_v - \hat{\mu_A})^2 / 2\hat{\sigma_A}^2}$$

[4]This constraint is particularly viable in biographic text, which tends to focus on the properties of a single individual.

[5]We use the hyperlinked phrases as potential values for all attributes except "Gender". For "Gender" we used pronouns as potential values ranked according to the their distance from the beginning of the page.

In the above equation, $posn_v$ is the absolute position ratio (position/length) and $\hat{\mu_A}, \hat{\sigma_A}^2$ are the sample mean and variance based on the sample of correct position ratios of attribute values in biographies with attribute A. Figure 2, for example, shows the positional distribution of the seed attribute values for deathdate, nationality and religion in Wikipedia articles, fit to a Gaussian distribution. Combining this empirically derived position model with a domain model[6] of acceptable attribute values is effective enough to serve as a stand-alone model.

Attribute	Best rank in seed set	P(Rank)
Birthplace	1	0.61
Birthdate	1	0.98
Deathdate	2	0.58
Gender	1	1.0
Occupation	1	0.70
Nationality	1	0.83
Religion	1	0.80

Table 1: Majority rank of the correct attribute value in the Wikipedia pages of the seed names used for learning relative ordering among attributes satisfying the domain model

6.1 Learning Relative Ordering in the Position-Based Model

In practice, for attributes such as birthdate, the first text pattern satisfying the domain model is often the correct answer for biographical articles. Deathdate also tends to occur near the beginning of the article, but almost always some point after the birthdate. This motivates a second, sequence-based position model based on the rank of the attribute values among other values in the domain of the attribute, as follows:

$$v^* = argmax_{v \in domain(A)} P(rank_v|A)$$

where $P(rank_v|A)$ is the fraction of biographies having attribute a with the correct value occuring at rank $rank_v$, where rank is measured according to the relative order in which the values belonging to the attribute domain occur from the beginning

[6]The domain model is the same as used in Section 4 and remains constant across all the models developed in this paper

of the article. We use the seed set to learn the relative positions between attributes, that is, in the Wikipedia pages of seed names what is the rank of the correct attribute.

Table 1 shows the most frequent rank of the correct attribute value and Figure 3 shows the distribution of the correct ranks for a sample of attributes. We can see that 61% of the time the first location mentioned in a biography is the individuals's birthplace, while 58% of the time the 2nd date in the article is the deathdate. Thus, "Deathdate" often appears as the second date in a Wikipedia page as expected. These empirical distributions for the correct rank provide a direct vehicle for scoring hypotheses, and the rows with "$rel.posn$" as the subscript in Table 4 shows the improvement in performance using the learned relative ordering. Averaging across different attributes, table 3 shows an absolute 11% average gain in accuracy of the position-sequence-based models relative to the improved Ravichandran and Hovy results achieved here.

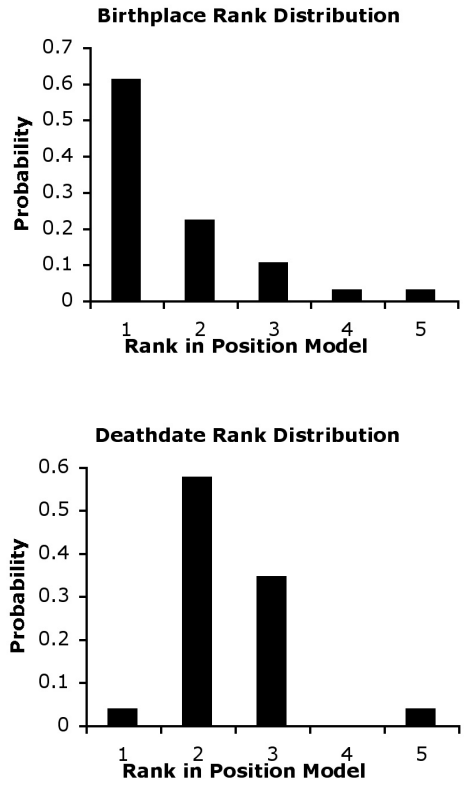

Figure 3: Empirical distribution of the relative position of the correct (seed) answers among all text phrases satisfying the domain model for "birthplace" and "death date".

7 Implicit Models

Some of the biographic attributes such as "Nationality", "Occupation" and "Religion" can be extracted successfully even when the answer is not directly mentioned in the biographic article. We present two such models for doing so in the following subsections:

7.1 Extracting Attributes Transitively using Neighboring Person-Names

Attributes such as "Occupation" are transitive in nature, that is, the people names appearing close to the target name will tend to have the same occupation as the target name. Based on this intution, we implemented a transitive model that predicts occupation based on consensus voting via the extracted occupations of neighboring names[7] as follows:

$$v^* = argmax_{v \in domain(A)} P(v|A, \mathcal{S}_{\text{neighbors}})$$

where,
$P(v|A, \mathcal{S}_{\text{neighbors}}) =$

$$\frac{\text{\# neighboring names with attrib value } v}{\text{\# of neighboring names in the article}}$$

The set of neighboring names is represented as $\mathcal{S}_{\text{neighbors}}$ and the best candidate value for an attribute A is chosen based on the the fraction of neighboring names having the same value for the respective attribute. We rank candidates according to this probability and the row labeled "trans" in Table 4 shows that this model helps in subsantially improving the recall of "Occupation" and "Religion", yielding a 7% and 3% average improvement in F-measure respectively, on top of the position model described in Section 6.

7.2 Latent Model based on Document-Wide Context Profiles

In addition to modeling cross-entity attribute transitively, attributes such as "Occupation" can also be modeled successfully using a document-wide context or topic model. For example, the distribution of words occuring in a biography

[7] We only use the neighboring names whose attribute value can be obtained from an encylopedic database. Furthermore, since we are dealing with biographic pages that talk about a single person, all other person-names mentioned in the article whose attributes are present in an encylopedia were considered for consensus voting

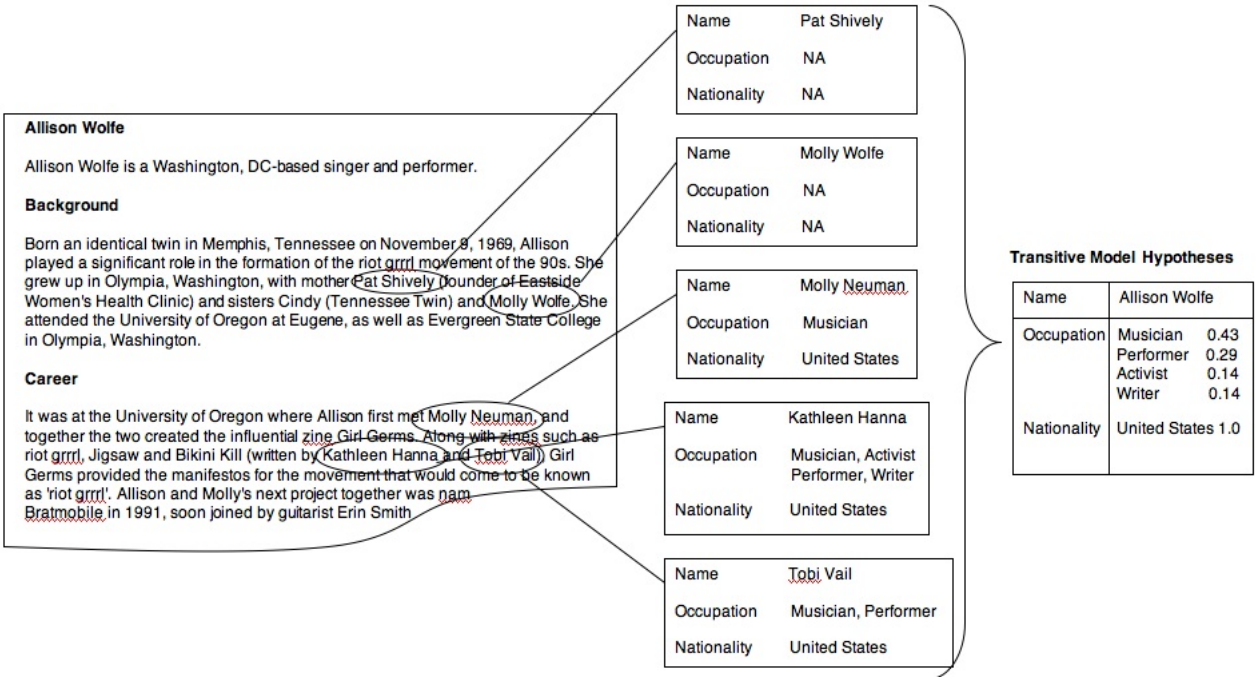

Figure 4: Illustration of modeling "occupation" and "nationality" transitively via consensus from attributes of neighboring names

of a politician would be different from that of a scientist. Thus, even if the occupation is not explicitly mentioned in the article, one can infer it using a bag-of-words topic profile learned from the seed examples.

Given a value v, for an attribute A, (for example v = "Politician" and A = "Occupation"), we learn a centroid weight vector:

$C_v = [w_{1,v}, w_{2,v}, ..., w_{n,v}]$ where,

$w_{t,v} = \frac{1}{N} tf_{t,v} \cdot log \frac{|A|}{|t \in A|}$

$tf_{t,v}$ is the frequency of word t in the articles of People having attribute $A = v$

$|A|$ is the total number of values of attribute A

$|t \in A|$ is the total number of values of attribute A, such that the articles of people having one of those values contain the term t

N is the total number of People in the seed set

Given a biography article of a test name and an attribute in question, we compute a similar word weight vector $C' = [w'_1, w'_2, ..., w'_n]$ for the test name and measure its cosine similarity to the centroid vector of each value of the given attribute. Thus, the best value a^* is chosen as:

$$v^* = argmax_v \frac{w'_1 \cdot w_{1,v} + w'_2 \cdot w_{2,v} + ... + w'_n \cdot w_{n,v}}{\sqrt{w'^2_1 + w'^2_2 + ... + w'^2_n} \sqrt{w^2_{1,v} + w^2_{2,v} + ... + w^2_{n,v}}}$$

Tables 3 and 4 show performance using the latent document-wide-context model. We see that this model by itself gives the top performance on "Occupation", outperforming the best alternative model by 9% absolute accuracy, indicating the usefulness of implicit attribute modeling via broad-context word frequencies.

This latent model can be further extended using the multilingual nature of Wikipedia. We take the corresponding German pages of the training names and model the German word distributions characterizing each seed occupation. Table 4 shows that English attribute classification can be successful using only the words in a parallel German article. For some attributes, the performance of latent model modeled via cross-language (noted as latentCL) is close to that of English suggesting potential future work by exploiting this multilingual dimension.

It is interesting to note that both the transitive model and the latent wide-context model do not rely on the actual "Occupation" being explicitly mentioned in the article, they still outperform ex-

Occupation	Weight Vector
	English
Physicist	<magnetic:32.7, electromagnetic:18.2, wire: 18.2, electricity: 17.7, optical:14.5, discovered:11.2>
Singer	<song:40, hits:30.5, hit:29.6, reggae:23.6, album:17.1, francis:15.2, music:13.8, recorded:13.6, ...>
Politician	<humphrey:367.4, soviet: 97.4, votes: 70.6, senate: 64.7, democratic: 57.2, kennedy: 55.9, ...>
Painter	<mural:40.0, diego:14.7, paint:14.5, fresco:10.9. paintings:10.9, museum of modern art:8.83, ...>
Auto racing	<renault:76.3, championship:32.7. schumacher:32.7, race:30.4, pole:29.1, driver:28.1 >
	German
Physicist	<faraday:25.4, chemie:7.3, vorlesungsserie:7.2, 1846:5.8, entdeckt:4.5, rotation:3.6 ...>
Singer	<song:16.22, jamaikanischen:11.77, platz:7.3, hit: 6.7, soloünstler:4.5, album:4.1, widmet:4.0, ...>
Politician	<konservativen:26.5, wahlkreis:26.5, romano:21.8, stimmen:18.6, gewählt:18.4, ...>
Painter	<rivera:32.7, malerin:7.6, wandgemälde:7.3, kunst:6.75, 1940:5.8, maler:5.1, auftrag:4.5, ...>
Auto racing	<team:29.4,mclaren:18.1,teamkollegen:18.1,sieg:11.7, meisterschaft:10.9, gegner:10.9, ...>

Table 2: Sample of occupation weight vectors in English and German learned using the latent model.

plicit pattern-based and position-based models. This implicit modeling also helps in improving the recall of less-often directly mentioned attributes such as a person's "Religion".

8 Model Combination

While the pattern-based, position-based, transitive and latent models are all stand-alone models, they can complement each other in combination as they provide relatively orthogonal sources of information. To combine these models, we perform a simple backoff-based combination for each attribute based on stand-alone model performance, and the rows with subscript "$combined$" in Tables 3 and 4 shows an average 14% absolute performance gain of the combined model relative to the improved Ravichandran and Hovy 2002 model.

9 Further Extensions: Reducing False Positives

Since the position-and-domain-based models will almost always posit an answer, one of the problems is the high number of false positives yielded by these algorithms. The following subsections introduce further extensions using interesting properties of biographic attributes to reduce the effect of false positives.

9.1 Using Inter-Attribute Correlations

One of the ways to filter false positives is by filtering empirically incompatible inter-attribute pairings. The motivation here is that the attributes are *not independent* of each other when modeled for the same individual. For example, P(Religion=Hindu | Nationality=India) is higher than P(Religion=Hindu | Nationality=France) and

Model	F_{score}	Acc truth pres
Ravichandran and Hovy, 2002	0.37	0.43
Improved RH02 Model	0.54	0.64
Position-Based Model	0.53	0.75
$\text{Combined}_{\text{above 3+trans+latent+cl}}$	0.59	0.78
Combined + Age Dist + Corr	**0.62** (+24%)	**0.80** (+37%)

Table 3: Average Performance of different models across all biographic attributes

similarly we can find positive and negative correlations among other attribute pairings. For implementation, we consider all possible 3-tuples of ("Nationality", "Birthplace", "Religion")[8] and search on NNDB for the presence of the tuple for any individual in the database (excluding the test data of course). As an agressive but effective filter, we filter the tuples for which no name in NNDB was found containing the candidate 3-tuples. The rows with label "combined+corr" in Table 4 and Table 3 shows substantial performaance gains using inter-attribute correlations, such as the 7% absolute average gain for Birthplace over the Section 8 combined models, and a 3% absolute gain for Nationality and Religion.

9.2 Using Age Distribution

Another way to filter out false positives is to consider distributions on meta-attributes, for example: while age is not explicitly extracted, we can use the fact that age is a function of two extracted attributes (<Deathyear>-<Birthyear>) and use the age distribution to filter out false positives for

[8]The test of joint-presence between these three attributes were used since they are strongly correlated

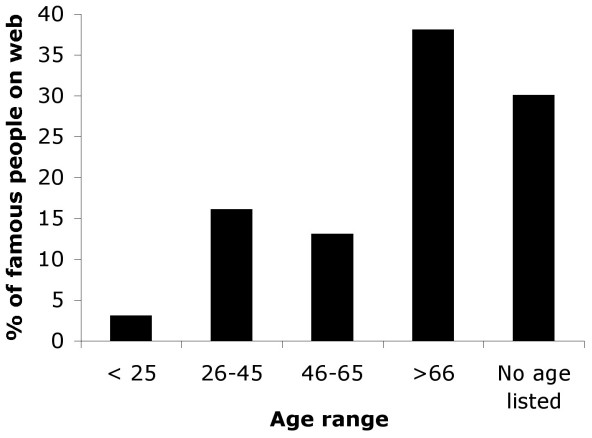

Figure 5: Age distribution of famous people on the web (from www.spock.com)

<Birthdate> and <Deathdate>. Based on the age distribution for famous people[9] on the web shown in Figure 5, we can bias against unusual candidate lifespans and filter out completely those outside the range of 25-100, as most of the probability mass is concentrated in this range. Rows with subscript "$comb + age\,dist$" in Table 4 shows the performance gains using this feature, yielding an average 5% absolute accuracy gain for Birthdate.

10 Conclusion

This paper has shown six successful novel approaches to biographic fact extraction using structural, transitive and latent properties of biographic data. We first showed an improvement to the standard Ravichandran and Hovy (2002) model utilizing untethered contextual pattern models, followed by a document position and sequence-based approach to attribute modeling.

Next we showed transitive models exploiting the tendency for individuals occurring together in an article to have related attribute values. We also showed how latent models of wide document context, both monolingually and translingually, can capture facts that are not stated directly in a text.

Each of these models provide substantial performance gain, and further performance gain is achived via classifier combination. We also showed how inter-attribution correlations can be modeled to filter unlikely attribute combinations, and how models of functions over attributes, such as deathdate-birthdate distributions, can further constrain the candidate space. These approaches collectively achieve 80% average accuracy on a test set of 7 biographic attribute types, yielding a 37% absolute accuracy gain relative to a standard algorithm on the same data.

[9]Since all the seed and test examples were used from nndb.com, we use the age distribution of famous people on the web: http://blog.spock.com/2008/02/08/age-distribution-of-people-on-the-web/

Attribute	Prec	P-Rec	F_{score}	Acc truth pres
$\text{Birthdate}_{\text{RH02}}$	0.86	0.38	0.53	0.88
$\text{Birthdate}_{\text{RH02}_{\text{imp}}}$	0.52	0.52	0.52	0.67
$\text{Birthdate}_{\text{rel. posn}}$	0.42	0.40	0.41	0.93
$\text{Birthdate}_{\text{combined}}$	0.58	0.58	0.58	0.95
$\text{Birthdate}_{\text{comb+age dist}}$	0.63	0.60	**0.61**	**1.00**
$\text{Deathdate}_{\text{RH02}}$	0.80	0.19	0.30	0.36
$\text{Deathdate}_{\text{RH02}_{\text{imp}}}$	0.50	0.49	0.49	0.59
$\text{Deathdate}_{\text{rel. posn}}$	0.46	0.44	0.45	**0.86**
$\text{Deathdate}_{\text{combined}}$	0.49	0.49	0.49	**0.86**
$\text{Deathdate}_{\text{comb+age dist}}$	0.51	0.49	**0.50**	**0.86**
$\text{Birthplace}_{\text{RH02}}$	0.42	0.38	0.40	0.42
$\text{Birthplace}_{\text{RH02}_{\text{imp}}}$	0.41	0.41	0.41	0.45
$\text{Birthplace}_{\text{rel. posn}}$	0.47	0.41	0.44	0.48
$\text{Birthplace}_{\text{combined}}$	0.44	0.44	0.44	0.48
$\text{Birthplace}_{\text{combined+corr}}$	0.53	0.50	**0.51**	**0.55**
$\text{Occupation}_{\text{RH02}}$	0.54	0.18	0.27	0.26
$\text{Occupation}_{\text{RH02}_{\text{imp}}}$	0.38	0.34	0.36	0.48
$\text{Occupation}_{\text{rel. posn}}$	0.48	0.35	0.40	0.50
$\text{Occupation}_{\text{trans}}$	0.49	0.46	0.47	0.50
$\text{Occupation}_{\text{latent}}$	0.48	0.48	**0.48**	**0.59**
$\text{Occupation}_{\text{latentCL}}$	0.48	0.48	0.48	0.54
$\text{Occupation}_{\text{combined}}$	0.48	0.48	**0.48**	**0.59**
$\text{Nationality}_{\text{RH02}}$	0.40	0.25	0.31	0.27
$\text{Nationality}_{\text{RH02}_{\text{imp}}}$	0.75	0.75	0.75	0.81
$\text{Nationality}_{\text{rel. posn}}$	0.73	0.72	0.71	0.78
$\text{Nationality}_{\text{trans}}$	0.51	0.48	0.49	0.49
$\text{Nationality}_{\text{latent}}$	0.56	0.56	0.56	0.56
$\text{Nationality}_{\text{latentCL}}$	0.55	0.48	0.51	0.48
$\text{Nationality}_{\text{combined}}$	0.75	0.75	0.75	0.81
$\text{Nationality}_{\text{comb+corr}}$	0.77	0.77	**0.77**	**0.84**
$\text{Gender}_{\text{RH02}}$	0.76	0.76	0.76	0.76
$\text{Gender}_{\text{RH02}_{\text{imp}}}$	0.99	0.99	0.99	0.99
$\text{Gender}_{\text{rel. posn}}$	1.00	1.00	**1.00**	**1.00**
$\text{Gender}_{\text{trans}}$	0.79	0.75	0.77	0.75
$\text{Gender}_{\text{latent}}$	0.82	0.82	0.82	0.82
$\text{Gender}_{\text{latentCL}}$	0.83	0.72	0.77	0.72
$\text{Gender}_{\text{combined}}$	1.00	1.00	**1.00**	**1.00**
$\text{Religion}_{\text{RH02}}$	0.02	0.02	0.04	0.06
$\text{Religion}_{\text{RH02}_{\text{imp}}}$	0.55	0.18	0.27	0.45
$\text{Religion}_{\text{rel. posn}}$	0.49	0.24	0.32	0.73
$\text{Religion}_{\text{trans}}$	0.38	0.33	0.35	0.48
$\text{Religion}_{\text{latent}}$	0.36	0.36	0.36	0.45
$\text{Religion}_{\text{latentCL}}$	0.30	0.26	0.28	0.22
$\text{Religion}_{\text{combined}}$	0.41	0.41	0.41	0.76
$\text{Religion}_{\text{combined+corr}}$	0.44	0.44	**0.44**	**0.79**

Table 4: Attribute-wise performance comparison of all the models across several biographic attributes.

References

E. Agichtein and L. Gravano. 2000. Snowball: extracting relations from large plain-text collections. *Proceedings of ICDL*, pages 85–94.

E. Alfonseca, P. Castells, M. Okumura, and M. Ruiz-Casado. 2006. A rote extractor with edit distance-based generalisation and multi-corpora precision calculation. *Proceedings of COLING-ACL*, pages 9–16.

J. Artiles, J. Gonzalo, and S. Sekine. 2007. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of SemEval*, pages 64–69.

S. Auer and J. Lehmann. 2007. What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. *Proceedings of ESWC*, pages 503–517.

A. Bagga and B. Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. *Proceedings of COLING-ACL*, pages 79–85.

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. *Proceedings of EACL*, pages 3–7.

J. Cowie, S. Nirenburg, and H. Molina-Salgado. 2000. Generating personal profiles. *The International Conference On MT And Multilingual NLP*.

S. Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. *Proceedings of EMNLP-CoNLL*, pages 708–716.

A. Culotta, A. McCallum, and J. Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. *Proceedings of HLT-NAACL*, pages 296–303.

E. Filatova and J. Prager. 2005. Tell me what you do and I'll tell you what you are: Learning occupation-related activities for biographies. *Proceedings of HLT-EMNLP*, pages 113–120.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING*, pages 539–545.

V. Jijkoun, M. de Rijke, and J. Mur. 2004. Information extraction for question answering: improving recall through syntactic patterns. *Proceedings of COLING*, page 1284.

G.S. Mann and D. Yarowsky. 2003. Unsupervised personal name disambiguation. In *Proceedings of CoNLL*, pages 33–40.

G.S. Mann and D. Yarowsky. 2005. Multi-field information extraction and cross-document fusion. In *Proceedings of ACL*, pages 483–490.

A. Nenkova and K. McKeown. 2003. References to named entities: a corpus study. *Proceedings of HLT-NAACL companion volume*, pages 70–72.

M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. 2006. Organizing and searching the World Wide Web of Facts Step one: The One-Million Fact Extraction Challenge. *Proceedings of AAAI*, pages 1400–1405.

D. Ravichandran and E. Hovy. 2002. Learning surface text patterns for a question answering system. *Proceedings of ACL*, pages 41–47.

Y. Ravin and Z. Kazi. 1999. Is Hillary Rodham Clinton the President? Disambiguating Names across Documents. *Proceedings of ACL*.

M. Remy. 2002. Wikipedia: The Free Encyclopedia. *Online Information Review Year*, 26(6).

E. Riloff. 1996. Automatically Generating Extraction Patterns from Untagged Text. *Proceedings of AAAI*, pages 1044–1049.

M. Ruiz-Casado, E. Alfonseca, and P. Castells. 2005. Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia. *Proceedings of NLDB 2005*.

M. Ruiz-Casado, E. Alfonseca, and P. Castells. 2006. From Wikipedia to semantic relationships: a semi-automated annotation approach. *Proceedings of ESWC*.

B. Schiffman, I. Mani, and K.J. Concepcion. 2001. Producing biographical summaries: combining linguistic knowledge with corpus statistics. *Proceedings of ACL*, pages 458–465.

M. Thelen and E. Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of EMNLP*, pages 14–21.

N. Wacholder, Y. Ravin, and M. Choi. 1997. Disambiguation of proper names in text. *Proceedings of ANLP*, pages 202–208.

C. Walker, S. Strassel, J. Medero, and K. Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium*.

R. Weischedel, J. Xu, and A. Licuanan. 2004. A Hybrid Approach to Answering Biographical Questions. *New Directions In Question Answering*, pages 59–70.

M. Wick, A. Culotta, and A. McCallum. 2006. Learning field compatibilities to extract database records from unstructured text. In *Proceedings of EMNLP*, pages 603–611.

L. Zhou, M. Ticrea, and E. Hovy. 2004. Multidocument biography summarization. *Proceedings of EMNLP*, pages 434–441.