Cascaded Information Synthesis for Timeline Construction

Gideon Mann
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
gideon.mann@gmail.com

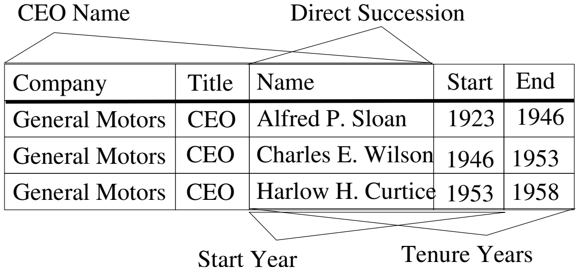
Abstract

This paper presents a method for incrementally constructing CEO succession timelines via a cascade of minimally supervised information synthesis components. Results from multiple documents are combined using a CRF fusion method. The system demonstrates that with minimal training on the target domain, the presence of redundant information allows for the synthesis of networks of interrelated facts from text.

1 Introduction

Single document information extraction of named entities and relationships has received much attention (e.g. MUC and ACE[1]). Relatively less explored is multi-document **information synthesis**, where information contained in separate documents within a large corpus is automatically extracted and fused to form networks of related facts. Timelines are an important example of such networks as they use temporal information to resolve ambiguities in extracted facts.

This paper presents a method for synthesizing CEO succession timelines from multiple documents without annotated data. The core of the method is an information synthesis component which performs document retrieval (Section 2.1), sentence extraction (Section 2.2), and cross-document fusion (Section 2.3).



Company	Title	Name	Start	End
General Motors	CEO	Alfred P. Sloan	1923	1946
General Motors	CEO	Charles E. Wilson	1946	1953
General Motors	CEO	Harlow H. Curtice	1953	1958

Figure 1: The goal of the system is a database filled with time-bounded CEO tenure facts. Separate information synthesis components fill different fields of the database incrementally.

A pipeline of these synthesis components is used to build a database of time-bounded CEO succession facts (Section 3). The pipeline produces a list of candidate CEOs (Section 3.1), direct transitions between CEOs (Section 3.2), CEO tenure midpoints (Section 3.3), and CEO start and end years (Section 3.4). System performance is given for each component as well as for the overall database (Section 5).

2 Information Synthesis via Retrieval, Extraction and Fusion

The goal of the system is to fill a relational database (e.g. Figure 1) in which each database record corresponds to a particular time-bounded fact. In the filled database, this collection of records constitutes a timeline. The system works incrementally to build up the entire database, leveraging a partially filled database at each step in the process. The filled fields in the database will be referred to as the **base fields**,

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/ and http://www.itl.nist.gov/iad/894.01/tests/ace/.

and the current cell being filled as the **target field**[2]. Each target field is also assigned a **target type** (e.g. PERSON) and words which belong to this type are marked during pre-processing stages. To find each target field, there is an information synthesis component composed of three stages: document retrieval, sentence extraction and cross-document fusion.

2.1 Document Retrieval

While very large corpora frequently contain redundant information, their use prohibits exhaustive application of complex information extraction methods. Evaluating a CRF on all documents on the Internet is infeasible. During document retrieval, a sub-corpus is selected from a larger corpus, allowing for deep processing of the documents that are most likely to contain the information of interest. The process of document retrieval is as follows: From the base fields, a **base query** is formulated. The query is issued to Google, which returns a ranked list of documents. The documents on this list are downloaded and preprocessed in series using the Penn Tokenizer (MacIntyre, 1995), a part-of-speech tagger (Florian and Ngai, 2001), and a Named Entity tagger[3].

2.2 Sentence Extraction

Once a set of documents has been gathered, sentences which contain the base fields and a candidate target[4] are selected, and a sentence extraction system is applied over the sentences. This paper uses Linear Chain Conditional Random Fields (CRFs) (Lafferty et al., 2001), an undirected conditional graphical model, to extract facts from sentences. A CRF yields a distribution $P(Y|X)$ of hidden labels Y for an observation sequence X:

$$P(Y|X) = \frac{1}{Z} \exp(\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, \mathbf{x}, t) \right)$$

Where Z is a normalization constant, t is a index into the linear chain, f_k is a binary feature function over the label y_t and its predecessor, the observation sequence, and the linear chain index t, and λ_k is the weight for a feature f_k.

During training, a base query is issued to Google for each training record, and the resulting corpus is then automatically annotated using the training base and target fields. This annotation is then used to train a CRF. The automatic annotation step marks all text which matches a base field or a target field. The models are trained with the assumption that base fields have been correctly filled. During training, in sentences where the target field isn't found, items of the target field type are labeled as spurious targets.

2.3 Cross-Document Fusion

After extraction is performed, it is necessary to fuse the targets to arrive at a consensus answer. Prior methods which use a CRF extractor for sentence level extraction have used **Viterbi frequency fusion (VFF)**, whereby the system chooses the fact mostly frequently extracted by the CRF Viterbi labeling sequence from all sentences in the corpus (Mann and Yarowsky, 2005).

This paper proposes using field confidence to fuse extracted facts. The field confidence is the probability of a word being assigned a certain label, summed over all other possible labels for the other words. If y_r is the label that indicates that a given word X_m is the target for a relationship r, then:

$$P(r(X_m)|X) = P(Y_m = y|X) = \sum_{Y':Y_m=y} P(Y'|X)$$

The field confidence can be efficiently computed using the Constrained Forward-Backward algorithm (Culotta and McCallum, 2004). From the field confidence the **maximum field confidence score (FCM)** for a given target over all sentences s can be computed :

$$C_{FCM}(X_m) = \max_s P(r(X^s_m)|X^s)$$

Alternatively, a **field confidence fusion (FCF)** score can be taken as sum over all sentences of the field confidence probability:

$$C_{FCF}(X_m) = \sum_s P(r(X^s_m)|X^s)$$

The fusion method described above is used for fusion of facts from one extractor. Sections 3.3 and 3.5

[2] In some cases, the target field is a derived field not explicitly present in the database.

[3] The Named Entity tagger is a CRF which is trained from the MUC-7 training corpora and achieves around 81% F-Measure on the MUC-7 test corpus.

[4] Candidate targets are identified by target type.

introduce additional methods for fusing facts from multiple extractors.

Cross-document fusion often requires some form of fact normalization, as the same target can be expressed in a variety of ways. For CEO succession timeline construction, the targets which exhibit the most variation are names. A simple name matcher was developed to perform this normalization, which uses a set of name-nickname pairs[5] to merge first names and optionally drops middle-initials.

3 Timeline Construction

For the problem of timeline construction, the information synthesis component described in the previous section is applied a number of times with different training data to synthesize different fields in the database (Figure 2). The pipeline is as follows:

1. Given a company, the system generates a set of CEOs for that company and the top candidate is picked (Section 3.1).

2. The direct succession model selects an adjacent CEO from the remaining candidates. (Section 3.2).

3. For each CEO in the pair, an estimated tenure midpoint (Section 3.3) and start and end tenure years (Section 3.4) are found. The pair order and the member start and end years are combined to arrive at a transition year estimate (Section 3.5).

3.1 CEO Name

The first step in the pipeline finds a list of candidate CEOs which are used for the remainder of the run. The base fields for the extractor are the company name and the title. The base query (e.g. "Boeing CEO") is issued to Google and the top 1000 documents are returned. The documents are marked with occurrences of the base field, the sentence extractor is applied over the sentences, and the extracted CEOs are then fused (Section 2.3). The system chooses the first CEO from the ranked list, and for further CEOs, the direct succession model is used.

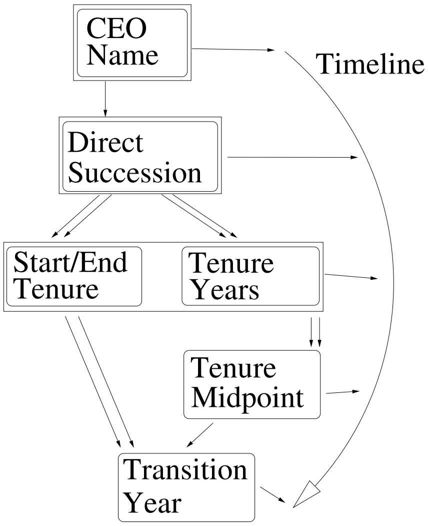


Figure 2: A series of information synthesis components builds up a CEO succession timeline. Later steps in the pipeline are less reliable and correspond to more fine-grained information.

3.2 Direct Succession

A second CEO is chosen from the ranked list using a **direct succession** model, which finds sentences that explicitly mention a succession relationship. The model is trained with two CEOs as a base, where the label sequence declares them to be in a particular order, and the target of extraction is the relative order of two given CEOs. For example, for the database shown in Figure 1, a training sentence might be marked:

> When/*pre* Alfred/*prev* Sloan/*prev* retired/*int* ,/*int* Charles/*next* Wilson/*next* replaced/*suf* him /*suf* ./*suf*

where the labels are *pre*fix, *prev*ious in relationship, *int*erstitial, *next* in relationship, and *suf*fix.

To choose a successor or predecessor for the top ranked CEO, each candidate ordered pair is separately searched for, and extraction and fusion are performed. Each base query is a pair of names (e.g. "Jeffrey-Immelt Jack-Welch"), and 100 documents from the Google returned list are returned for each pair.

3.3 Tenure Midpoint Estimation

For **tenure midpoint** estimation, the base is a CEO name and the target is the midpoint of the CEO's

[5] collected by P. Driscoll.

tenure in that position. For this component and subsequent components, each base query is a CEO name and company (e.g. "GE Jack-Welch"), and 100 documents are retrieved. The tenure midpoint is estimated by extracting years for which the CEO was in office and taking a weighted sum over the list of years. To build the sentence extractor to extract the years in which the CEO was in office, these tenure years are marked in the training corpus. Although the confidence estimates for particular tenure years are often noisy, the estimated tenure midpoints provide a second source of information about the relative ordering between two people.

3.4 Start/End Tenure Year

For start and end tenure years, the base fields are CEOs and the targets are the desired year. Start and end years are the least reliable information from all of the different methods, because they do not appear frequently in text, whereas all of the prior methods are able to use corpus redundancies to boost accuracy.

3.5 Transition Year Estimation

Underlying data dependencies are useful in increasing the accuracy of the exact start/end years. Knowing the information `succeed(A,B)`, then it must be true that `end(A,X)` and `start(B,X)`. Two methods for using this information are considered. In the first, the end tenure year for A is thrown out and replaced by `start(B,x)`, as start year prediction is known to be higher confidence. In the second method, a linear combination is used to provide a new estimate of the transition year. Given $C^E_A(X)$, the confidence for the end year of the predecessor A, and $C^S_B(X)$, the confidence for the start year of the successor B, and the estimate for $C^T_{AB}(X)$, the confidence for a transition year X is:

$$C^T_{AB}(X) = C^E_A(X) \times C^S_B(X)$$

4 Example Pipeline Run

An example of the pipeline is shown for the company Gannett. Table 1 shows an example top ten list for Gannett CEO extraction. The top ten list contains 5/6 of the total possible Gannett CEOs, where incorrect candidates are primarily heads of other divisions within Gannett:. The top choice, Douglas McCorkindale is entered into the database.

Name	Confidence
Douglas H. McCorkindale	0.181
Craig A. Dubow	0.092
Allen H. Neuharth	0.068
Cecil L. Walker	0.065
John J. Curley	0.046
Frank Gannett	0.042
Roger L. Ogden	0.01
Ken Tonning	0.006
Craig Moon	0.006
Mimi Feller	0.005

Table 1: CEOs extracted for Gannett. Correct CEOs in **bold**. Douglas McCorkindale, the top ranked candidate, was CEO of Gannett from 2000-2005.

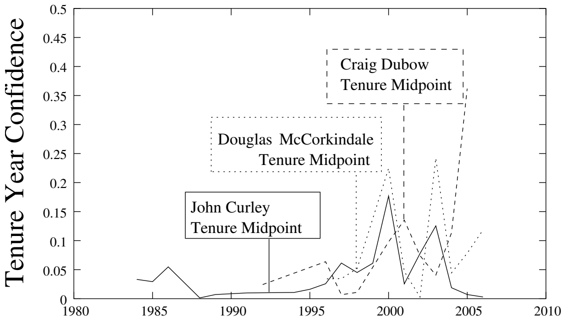
Proposed CEO order (A,B) X = McCorkindale	Confidence
X, Craig Dubow	5.193
X, John Curley	4.460
John Curley, X	3.835
Craig Dubow, X	3.113
Craig Moon, X	1.034
X, Craig Moon	1.007
X, Cecil Walker	0.949
Frank Gannett, X	0.927
Cecil Walker, X	0.854
X, Frank Gannett	0.124

Table 2: FCF scores for pairs in Direct Succession Model. Correct relative orderings in **bold**. (McCorkindale, Dubow) is the highest ranking proposed pair.

The system then picks a second CEO from the list using the direct succession model. Table 2 lists the field confidence fusion scores associated with various ordered pairs. The top pair, (Douglas McCorkindale, Craig Dubow), is entered into the database.

The tenure years for McCorkindale and Dubow are then extracted and tenure midpoints are estimated. McCorkindale (2000-2005) has an estimated tenure midpoint of 1998. Dubow (2005-) has an estimated tenure midpoint of 2001. Given these tenure midpoints, the ordering proposed in the previous step is confirmed, with McCorkindale preceding Dubow.

The start and end years for McCorkindale and



Company	Title	Name	Start	End
Gannett	CEO	John Curley	x	2000
Gannett	CEO	Douglas McCorkindale	2000	2005
Gannett	CEO	Craig Dubow	2005	x

Figure 3: Graph of CEO tenure year confidence for Gannett. The three most recent CEOs have been correctly identified. Below the graph is the extracted database.

Dubow are extracted as well. The estimated span points for McCorkindale are (2000-2003), with the correct year, 2005, ranked 10th. For Dubow, the estimated start and end years are both 2005. In this case the re-estimated transition year from fusion yields the year 2005, which is correct. If the system were to perform another iteration, CEO John Curley would be found, producing the timeline depicted in Figure 3.

5 Detailed Experimental Results

The database used for the experiments reported in the following section took a sample of 18 companies from the Fortune 500 list[6]. For each company, the author used the Internet to find ground truth of the entire CEO history for the company. Of the companies, six were randomly chosen and selected for training (Anheuser-Busch, Hewlett-Packard, Lennar, McGraw-Hill, Pfizer, and Raytheon), four were used as a development set (Boeing, Heinz, Staples, and Textron), and eight were used for testing (Gannett, General Electric, General Motors, HomeDepot, IBM, Kroger, Sears, and UPS). Altogether, there were 98 database records for the companies, with 21 training records, 16 development records, and 61 testing records.

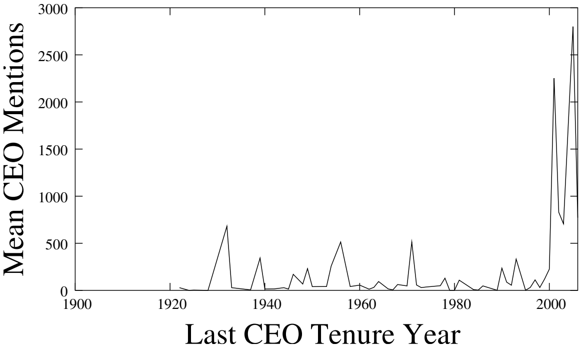


Figure 4: CEO mentions over time. Much more information is available for recent CEOs than for CEOs from earlier decades.

One peculiarity of the data is that CEOs from earlier decades have very few mentions on the web as compared with more recent CEOs (see Figure 4). Many of the missing CEOs in the CEO name extraction step never appeared on the web. For example, for Fowler McConnell, Sears CEO from 1958 to 1960, the query "Fowler-McConnell Sears" returns exactly 2 documents.

5.1 CEO Name

The precision and recall of the returned ranked lists can be calculated for the CEO name component. Figure 5 graphs precision and recall for CEO extraction on the 8 test set companies using the Viterbi frequency fusion (VFF), field confidence maximum (FCM), and field confidence fusion (FCF). Precision near the top of the ranked list is quite high – more than 90% of the returned top 2 CEOs are correct. Further, recall never reaches more than 70%. As previously mentioned, this is primarily due to sparse data for CEOs in earlier decades, before periodicals were published widely on the Internet. Finally, FCF performs slightly better than VFF and FCM, particularly towards the top of the ranked list, which is the most crucial.

5.2 Synthesis of Temporal Information

Once the CEO name list has been extracted and the top CEO candidate selected as a future base fact, the system begins to fill the database with temporal information. The system finds a high confidence direct succession pair and uses tenure midpoint as a re-estimation procedure. It then finds start and end

[6] `http://www.fortune.com/fortune/fortune500`

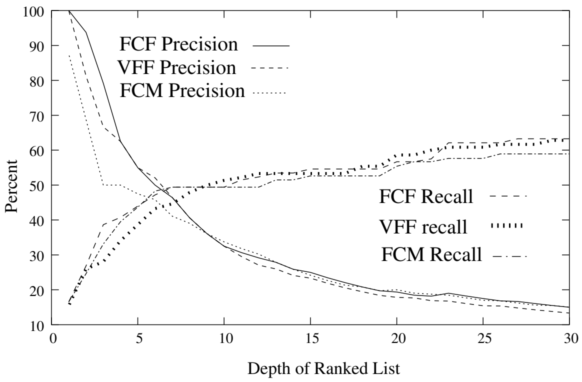


Figure 5: Precision/Recall of CEO name lists averaged over the test companies.

	VFF		FCF	
Component	Acc.	MRR	Acc.	MRR
Direct Succession	62.5%	79%	**75%**	**85.4%**
Tenure Midpoint	50%	N/A	**62.5%**	N/A
Start Year	37.5%	40.6%	**50%**	**55.7%**
End Year	**12.5%**	12.5%	6.25%	**15.1%**

Table 3: Synthesis of temporal information for the 8 companies and 16 people in the set. The accuracy and mean reciprocal rank of two fusion methods are compared. FCF performs better than VFF on most tasks.

years and uses the start years and ordering to estimate the transition year. For the direct succession and tenure midpoint reordering, there are 8 pairs to be evaluated, one for each company. For the start and end years, there are 16 people (one pair of CEOs for each company).

Table 3 and shows performance for each target field for FCF and VFF. The results suggest that Viterbi frequency fusion is typically less effective than field confidence fusion. Only in one case (end year accuracy) did VFF perform better than FCF. The Direct Succession model performed slightly better than the Tenure Midpoint model for ordering. The tenure midpoint model may still be useful to provide corroborating evidence for the direct succession model.

Table 4 shows the relative performance of FCM and FCF, where FCM is the single document best field confidence score. On the whole, FCF outperforms FCM, though for end tenure year synthesis FCM is more successful.

Component	FCM	FCF
Direct Succession	63.5%	**75%**
Start Year	43.8%	**50%**
End Year	**20%**	6.25%

Table 4: Comparison of field confidence fusion (FCF) and maximum field confidence (FCM) methods.

The performance for start and end year synthesis was considerably lower than for the previous steps, primarily because there are few data redundancies to exploit. End year results were poor in part because for the CEOs still in office, all answers were graded as incorrect. Additionally, the fact that start years are more common in the corpus than end years caused errors in which the start year was returned as the end year.

For re-estimating transition years, both picking the start year of the succeeding member of the pair and picking a linear combination of the confidences were used. For the pairs, 50% of the transition years are predicted correctly using the start year of the next candidate, and 25% using linear interpolation.

5.3 Overall Database Accuracy

In total, there were 61 testing database records, with 3 fields per record to be discovered (CEO name, start year, and end year). The derived field of relative ordering is not graded, though it has precise information which would be useful for seeding future information synthesis systems.

The initial system returns only 16 records, and so has a recall of 26%. The low recall is due primarily to the lack of the information on the web for CEOs from decades before 1990. For the 48 targets fields recovered by the system, 27 (56%) are correct, where most of the errors are for incorrectly extracted end years. This grading criterion is strict as it penalizes cases where the found CEO is still in office. With end years removed for CEOs in office, the performance in correctly filled fields is 67%.

	Query Once		Query Many
Component	100 pgs	1000 pgs	100 pgs
Direct Succession	50%	37.5%	**75%**
Tenure Midpoint	37.5%	**62.5%**	**62.5%**
Start Year	37.5	37.5%	**50%**
End Year	**31.25**	18.75%	6%

Table 5: Allowing each information synthesis component to retrieve its own corpus yields higher performance than fixing the entire corpus at the start.

5.4 Multiple IR Queries

There are multiple separate corpora downloaded for different base queries: one for company names and title in the CEO name component (up to 1000 pages each), distinct corpora for each ordered pair of CEOs in the direct succession component (100 pages each), and another set of distinct corpora for individual CEOs for tenure years as well as start and end tenure points (100 pages each). In order to test the impact of these distinct information retrieval steps, results were compared with those generated from a system using only the initial corpus.

Table 5 shows the relative performance for single document retrieval step of 100 or 1000 documents as opposed to incremental retrieval steps of 100 documents as the database is partially filled. Performing multiple queries appears to have an edge over downloading one corpus, especially in the direct succession and tenure midpoint estimation steps. For end tenure year the larger corpus lead to better performance, which suggests a possible inefficiency in the retrieval set for CEO end tenure year.

6 Related Work

The term "information synthesis" has been used by (Blake and Pratt, 2002) and (Blake, 2005) to describe a human-computer collaborative process of retrieval, extraction, and analysis of research literature. Amigo et al. (2004) use the term information synthesis for "topic oriented, multi-document summarization".

There has been relatively little work on extraction of temporal facts. There is related work in temporal summarization by sentence selection to create timelines (Allan et al., 2001; Chieu and Lee, 2004), and TIMEX extraction and resolution (Mani and Wilson, 2000). Pustejovsky et al. (2003) describe a language for annotating time events, but does not provide a way to extract this information.

Brin (1998), Agichtein and Gravano (2000), and Ravichandran and Hovy (2002) present related methods for training an information extraction system by example facts. Alternative types of training can be found in (Riloff, 1996), which trains from texts marked relevant and irrelevant, and (Etzioni et al., 2005), which trains from single example patterns such as "*actor* starred in *film*".

Fusion of extracted facts is a relatively new area of investigation. Prior work in the area includes (Skounakis and Craven, 2003) and (Downey et al., 2005) which present models for information fusion for facts extracted by classifiers. Closely related are (Finkel et al., 2005) and (Sutton and McCallum, 2004) which present methods for the joint labeling of named entities in text using graphical models for single document extraction. Mann and Yarowsky (2005) introduce two simple methods for fact fusion for sequence models.

The problem of management succession has been studied in the context of the MUC-6 evaluation (Grishman and Sundheim, 1996), which included an evaluation of extraction of management succession events from single document. Most of the systems developed for this task were hand-crafted, knowledge engineered systems. Notable exceptions are (Soderland, 1999), which learned a set of regular expressions, and (Chieu and Ng, 2002), which used a log-linear classifier. Both systems extract non-temporal succession events and find the company, the position, the previous position holder, and the successor. For that task, Chieu and Ng (2002) reports results of 60% F-measure for multi-slot management succession extraction from a single document but does not extract start and end years. These improved results can be attributed to the presence of labeled training data (6915 annotated instances), a more homogeneous corpus made up of newswire, and matched training and test data. Additionally, the system doesn't evaluate on start and end year extraction, the targets found in this paper to be the most difficult.

7 Conclusion

This paper presented a system for synthesizing time-bounded facts from large corpora for timeline construction. This is a novel information analysis task which is made possible by minimally supervised training of sentence extractors, redundant corpora that compensate for noisy extraction, and dependencies between related facts. Incremental construction of databases by linked information synthesis components allows for the gradual aggregation of semantic networks of facts, and the data synthesized in this paper could serve as input to yet another processing step.

An information synthesis component was presented which retrieved relevant documents, extracted facts from sentences, and fused the resulting facts. Field confidence fusion was shown to be an effective method for cross-document fusion. Research into additional synthesis components which rely on information not present in a single sentence is a promising area of future work.

The resultant timelines provide recent CEO succession information including relative order, start years, and end years. These temporal attributes are fundamental properties of time-bounded facts and may be used for related synthesis tasks.

Please contact the author for access to the training, development, and test database of CEO tenure information.

8 Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

References

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

J. Allan, R. Gupta, and V. Khandelwal. 2001. Temporal summaries of news topics. In *SIGIR*, pages 10–18.

E. Amigo, J. Gonzalo, V. Peinado, A. Penas, and F. Verdejo. 2004. An empirical study of information synthesis task. In *ACL*.

C. Blake and W. Pratt. 2002. Collaborative information synthesis. In *ASIST*.

C. Blake. 2005. Information synthesis: A new approach to explore secondary information in scientific literature. In *JCDL*.

S. Brin. 1998. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, pages 172–183.

H. L. Chieu and Y. K. Lee. 2004. Query based event extraction along a timeline. In *SIGIR*, pages 425–432.

H. L. Chieu and H. T. Ng. 2002. A maximum entropy approach to information extraction from semi-structured and free text. In *AAAI*, pages 786–791.

A. Culotta and A. McCallum. 2004. Confidence estimation for information extraction. In *Proceedings of HLT-NAACL*.

D. Downey, O. Etzioni, and S. Soderland. 2005. A probabilistic model of redundancy in informaion extraction. In *IJCAI*.

O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

J. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.

R. Florian and G. Ngai. 2001. Multidimensional transformation-based learning. In *Proceedings of CoNLL*.

R. Grishman and B. Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of COLING*.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

R. MacIntyre. 1995. Penn tokenizer.

I. Mani and G. Wilson. 2000. Robust temporal processing of news. In *Proceedings of ACL*, pages 69–76.

G. Mann and D. Yarowsky. 2005. Multi-field information extraction and cross-document fusion. In *Proceedings of ACL*.

J. Pustejovsky, J. Castao, R. Ingria, R. Saur, R. Gaizauskas, A. Setzer, and G. Katz. 2003. Timeml: Robust specification of event and temporal expressions in text. In *IWCS-5 Fifth International Workshop on Computational Semantics*.

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

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

M. Skounakis and M. Craven. 2003. Evidence combination in biomedical natural language processing. In *SIGKDD Workshop on Data Mining in Bioinformatics*.

Stephen Soderland. 1999. Learning Information Extraction Rules for Semi-structured and Free Text. *Machine Learning*.

C. Sutton and A. McCallum. 2004. Collective segmentation and labeling of distant entities in information extraction. In *ICML workshop on Statistical Relational Learning*.