Why Read if You Can Scan?
Trigger Scoping Strategy for Biographical Fact Extraction

Dian Yu and Heng Ji
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY, USA
{yud2,jih}@rpi.edu

Sujian Li
Peking University
Key Laboratory of
Computational Linguistics
Beijing, China
lisujian@pku.edu.cn

Chin-Yew Lin
Microsoft Research Asia
Beijing, China
cyl@microsoft.com

Abstract

The rapid growth of information sources brings a unique challenge to biographical information extraction: how to find specific facts without having to read all the words. An effective solution is to follow the human scanning strategy which keeps a specific keyword in mind and searches within a specific scope. In this paper, we mimic a scanning process to extract biographical facts. We use event and relation *triggers* as keywords, identify their scopes and apply type constraints to extract answers within the scope of a trigger. Experiments demonstrate that our approach outperforms state-of-the-art methods up to 26% absolute gain in F-score without using any syntactic analysis or external knowledge bases.

1 Introduction

Extracting biographical information is an important task because it can help readers understand an ongoing event more easily by providing the background biographical information of participants in this event. In fact, this task has been part of the Text Analysis Conference (TAC) - Knowledge Base Population (KBP) Slot Filling (SF) Track (Ji et al., 2010; Ji et al., 2011; Surdeanu, 2013; Surdeanu and Ji, 2014) for years.

Overall, state-of-the-art research still needs improvement. A typical approach is based on patterns which include triggers (*e.g.*, (Sun et al., 2011; Li et al., 2012)). Here *trigger* is defined as the smallest extent of a text which most clearly expresses an event occurrence or indicates a relation type. High-quality patterns yield quite high precision but relatively low recall. In addition, it's relatively expensive to maintain and update a set of extraction patterns.

Furthermore, we carefully investigated the TAC-KBP SF 2012 ground truth corpus and find that 94.36% of the biographical facts are mentioned in a sentence containing indicative fact-specific triggers. For example, *born* is a trigger for extracting birth-related facts. Triggers are crucial in predicting the type of facts (Aguilar et al., 2014). However, most previous studies only focused on using triggers to create more patterns (*e.g.*, (Li et al., 2013)). Therefore the critical problem is how to make the most of triggers in biographical fact extraction?

We observe that people tend to scan a document when they want to quickly find a biographical fact within limited time. According to Douglas and Frazier (2001), *scanning* is a strategy for quickly finding specific information (keywords or ideas) in a text while ignoring its broader meaning. Scanning involves skipping words, but the emphasis is that the reader knows what to look for and rapidly scans until words are found and closer reading can occur (Phipps, 1983).

There are five steps in implementing scanning strategy according to Arnold (1999):

1. Keep in mind what you are searching for.
2. Anticipate in what form the information is likely to appear – number, proper nouns, etc.
3. Analyze the organization of the content before starting to scan.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1203–1208,
Denver, Colorado, May 31 – June 5, 2015.

4. Let your eyes run rapidly over several lines of print at a time.

5. When you find the sentence that has the information you seek, read the entire sentence.

Educators have verified that scanning is an effective strategy in enhancing reading comprehension (Motallebzadeh and Mamdoohi, 2011). There are two important aspects in the scanning strategy: keywords and their corresponding scopes. For biographical fact extraction, triggers can easily act as the keywords used by human during scanning and thus we focus on identifying the scopes of triggers.

Given a sentence that contains one or more triggers, we define *trigger scope* as the shortest fragment that is related to a trigger. Based on our observation, each fact-specific trigger has its own scope and its corresponding facts seldom appear outside of its scope. In the following sentence, if we can identify the scope of *graduated*, a trigger for education-related facts, we can skip the rest of the sentence after *1965* even though *Chesterbrook Academy* is an educational organization.

*She [<**graduated**> from **Barnard** in 1965] and soon began teaching English at **Chesterbrook Academy** in Pennsylvania.*[1]

In this paper, we study the effect of triggers by learning their linguistic scopes at the sentence level and apply this strategy to extract 11 types of biographical facts, namely, *birth date, death date, birth place, death place, residence place, education, parents, spouse, children, siblings* and *other family* as described in the KBP SF task.

We design our extraction process following the scanning steps corresponding to Arnold's scanning theory.

1. Let the computer know the query and the fact type to be extracted.

2. Let the computer know what form or entity type the candidate answer is likely to appear – person, organization, phrase, time, etc.

3. Locate all the triggers of the given fact type and recognize their respective scopes.

4. Within each scope, extract candidate answers which satisfy the entity type constraint in 2.

The contributions of our paper are as follows.

- We are the first to study the application of trigger scoping in biographical fact extraction.

- Our approach does not rely on any external knowledge bases for training or manually created fact-specific rules, and yet dramatically advances state-of-the-art.

2 Approach

In this section, we present the detailed approach of applying trigger scoping to biographical fact extraction. In Section 2.1, we first introduce the annotation methods of constructing the gold-standard dataset for evaluating scope identification. We use the sentence in Figure 1 as our illustrative example.

> *Paul Francis Conrad* and his [twin <**brother**>, James], were [<**born**> in Cedar Rapids, Iowa, on June 27, 1924], [<**sons**> of Robert H. Conrad and Florence Lawler Conrad].

Figure 1: Trigger and scope annotation example.

2.1 Trigger and Scope Annotation

2.1.1 Basic issues

In a text, the sentences containing biographical facts (*e.g.*, birth, death, family, residence or education) are considered for annotation. We discard a sentence if it expresses a biographical fact without surface cues.

During annotation, triggers are marked by angle brackets (*e.g.*, <*resident*>), and the scope boundaries of a trigger are denoted by square brackets as shown in Figure 1.

2.1.2 Trigger Tagging

We mined fact-specific trigger lists from existing patterns (Chen et al., 2010; Min et al., 2012; Li et al., 2012) and ground truth sentences from KBP 2012

[1] The scope is marked with [] and the trigger is marked with <>.

SF corpus. In our experiment, we use 343 triggers and 38 triggers on average for each fact type[2].

We examine all the sentences containing any possible triggers. The presence of a word in one trigger list does not necessarily mean that the sentence contains an event or a relation. For instance, the second *child* in the following sentence is part of an organization's name.

*He and his wife, Ann McGarry Buchwald moved to Washington in 1963 with their [<**child**>], who was adopted from orphanages and [<**child**> welfare agencies] in Ireland, Spain and France.*

We also keep such sentences and annotate their trigger scopes without distinction.

Note that we only mark the syntactic head of a trigger phrase. For example, we mark *child* for the noun phrase *the second child*.

2.1.3 Scope Tagging

During the scope annotation, we first include the trigger within its own scope and then mark its left and right boundaries. Usually the left boundary is the trigger itself.

When there are multiple triggers in the same sentence, we annotate each trigger's scope separately since it is possible that the scopes of different triggers are overlapped or nested as shown in the following instance (the scope of *daughters* covers the scope of *wife*):

*Pavarotti had three [<**daughters**> with his first wife, Lorenza, Cristina and Giuliana; and one, Alice, with his second wife].*

*Pavarotti had three daughters with his first [<**wife**>], Lorenza, Cristina and Giuliana; and one, Alice, with his second [<**wife**>].*

The scope of a word is not transitive. In the phrase *"his [<**son**>'s home] in Washington"*, *home* is within *son*'s scope and *in Washington* is within *home*'s scope, however, the last prepositional phrase is outside of *son*'s scope.

2.2 Scope Identification

We will introduce two methods for identifying trigger scopes.

[2]The trigger lists are publicly available for research purposes at: http://nlp.cs.rpi.edu/data/triggers.zip

2.2.1 Rule-based Method

This method is used to investigate the performance of trigger scoping strategy when we do not have any labeled data. We use trigger as the left scope boundary. A verb or trigger with other fact types is regarded as the right boundary.

The rule-based scoping result of the walk-through example is as follows:

> *Paul Francis Conrad* and his twin [<**brother**>, James, were] [<**born**> in Cedar Rapids, Iowa, on June 27, 1924,] [<**sons**> of Robert H. Conrad and Florence Lawler Conrad.]

2.2.2 Supervised Classification

Alternatively we regard scope identification as a classification task. For each detected trigger, scope identification is performed as a binary classification of each token in the sentence as to whether it is within or outside of a trigger's scope.

We apply the Stanford CoreNLP toolkit (Manning et al., 2014) to annotate part-of-speech tags and names in each document. We design the following features to train a classifier.

- Position: The feature takes value 1 if the word appears before the trigger, and 0 otherwise.
- Distance: The distance (in words) between the word and the trigger.
- POS: POS tags of the word and the trigger.
- Name Entity: The name entity type of the word.
- Interrupt: The feature takes value 1 if there is a verb or a trigger with other fact type between the trigger and the word, and 0 otherwise. Verbs and triggers with other fact types can effectively change the current topic or continue in another way.

Note that the trained classifier can make predictions that result in nonconsecutive blocks of scope tokens. In this case, we aggregate the labels of all the words of an entity to assign a global label, which means that we assign the entity the majority label of the words it contains.

Fact Type	Recall (%)			Precision (%)			F-score (%)		
	1	2	3	1	2	3	1	2	3
per:place_of_birth	59.4	88.2	88.2	76.0	87.0	88.2	66.7	87.6	**88.2**
per:date_of_birth	59.1	94.4	100.0	100.0	94.4	100.0	74.3	94.4	**100.0**
per:place_of_death	55.4	92.4	86.1	86.1	58.9	63.6	67.4	71.9	**73.1**
per:date_of_death	46.4	98.2	96.5	81.3	48.3	53.4	59.1	64.7	**68.8**
per:place_of_residence	60.0	68.9	68.9	40.4	64.2	61.3	48.3	**66.5**	64.9
per:school_attended	54.3	65.8	68.4	86.4	67.6	76.5	66.7	66.7	**72.2**
per:parents	41.9	75.7	73.0	68.4	31.8	50.0	52.0	44.8	**59.3**
per:sibling	50.0	76.2	76.2	61.5	59.3	55.2	55.2	**66.7**	64.0
per:spouse	36.0	63.3	81.7	78.3	54.3	49.5	49.3	58.5	**61.6**
per:children	39.5	61.8	76.4	73.2	58.5	71.6	51.3	60.1	**73.9**
per:other_family	23.1	66.7	71.4	75.0	53.9	53.6	35.3	59.6	**61.2**
overall	47.7	77.4	80.6	75.1	61.7	65.7	56.9	67.4	**71.6**

Table 1: performance on KBP 2013 (1:state-of-the-art; 2:rule-based; 3: SVMs).

2.3 Biographical Fact Extraction

For each relevant document of a given query, we use Stanford CoreNLP to find the coreferential mentions of the query and then return all the sentences which contain at least one query entity mention. For each trigger in a sentence, we extract the entities which satisfy fact-specific constraints within its scope. As shown in Figure 1, *brother* is the trigger for *per:siblings* and the candidate fact should be a person name. Thus we return all the person names (*e.g.*, *James*) within *brother*'s scope as the query *Paul*'s siblings.

3 Experiments and Discussion

3.1 Data

We use the KBP 2012 and 2013 SF corpora as the development and testing data sets respectively. There are 50 person queries each year.

From the KBP 2012 SF corpus, we annotated 2,806 sentences in formal writing from news reports as the gold-standard trigger scoping data set. We randomly partitioned the labeled data and performed ten-fold cross-validation using LIBSVM toolkit (Chang and Lin, 2011). We employ the classification model trained from all the labeled sentences to classify tokens in the unlabeled sentences.

3.2 Results

3.2.1 Scope Identification

The scope identification evaluation results of the rule-based method and the SVMs with the RBF kernel are presented in Table 2. We can see that the supervised classification method performs better since it incorporates the weights of different features rather than simply applying hard constraints. In addition, it allows the answers to appear before a trigger as shown in the following sentence. Our rule-based method fails to extract *Fred* since it appears before the trigger *married*:

*She was a part of a group of black intellectuals who included philosopher and poet [**Fred Clifton**, whom she <**married**> in 1958].*

Fact Group	Accuracy (%)		F-score (%)	
	Rule	SVMs	Rule	SVMs
Birth	85.97	**96.66**	80.01	**94.21**
Death	92.31	**94.56**	82.16	**89.01**
Residence	90.67	**95.67**	76.11	**83.25**
Family	92.49	**94.11**	75.30	**77.31**
Education	91.51	**93.87**	88.46	**90.65**

Table 2: Scope identification results.

3.2.2 Biographical Fact Extraction

The fact extraction results in Table 1 demonstrate our trigger scoping strategy can outperform state-of-the-art methods. For a certain fact type, we choose the SF system which has the best performance for comparison. Specifically, we compare with two successful approaches: (1) the combination of distant supervision and rules (*e.g.*, (Grishman, 2013; Roth et al., 2013)); (2) patterns based on dependency paths (*e.g.*, (Li et al., 2013; Yu et al., 2013)).

The advantage of our method lies in trigger-driven

exploration. The positions of facts in the sentence can be very flexible and therefore difficult to be captured using a limited number of patterns. For example, the patterns in table 2[3] fail to extract *James* in Figure 1. However, the ways in which we express the trigger and words it dominated tend to be relatively fixed. For example, all the following patterns contain a fact-specific trigger and also facts usually appear within its scope.

PER:SIBLING
[Q] poss^{-1} **brother** appos **[A]**
[Q] appos^{-1} **brother** appos **[A]**
[Q] appos **brother** appos-1 **[A]**
[Q] nsubjpass^{-1} survived agent **brother** appos **[A]**
[Q] poss^{-1} **sister** appos **[A]**
[Q] appos^{-1} **sister** appos **[A]**
[Q] appos **sister** appos^{-1} **[A]**
[Q] nsubjpass^{-1} survived agent **sister** appos **[A]**

Table 3: Patterns used for extracting sibling facts (Li et al., 2013). Q: Query, A: Answer.

The limitation of our method is that we assume a sentence centers around only one person thus every biographical fact mentioned should be related to the centroid person. For example, our method mistakenly extracted *February* as the death-date fact for both *Reina* and *Orlando* in the following case.

*Also at the mass was **Reina Tamayo**, the mother of **Orlando Zapata**, who [<**died**> in February] after an 85-day hunger strike to protest the fate of political prisoners here.*

In order to solve this problem, we need to further analyze the relation between the query entity mention and the trigger so that we can identify *Orlando Zapata* is irrelevant to the death-related fact.

4 Related Work

Previous successful approaches to construct the biographical knowledge base are relatively expensive: Distant Supervision (Surdeanu et al., 2010) relies upon external knowledge bases and it is time-consuming to manually write or edit patterns (Sun et al., 2011; Li et al., 2012). The main impact of our trigger scoping strategy is to narrow down the text span of searching for facts, from sentence-level to fragment-level. We only focus on analyzing the content which is likely to contain an answer.

Our trigger scoping method is also partially inspired from the negation scope detection work (*e.g.*, (Szarvas et al., 2008; Elkin et al., 2005; Chapman et al., 2001; Morante and Daelemans, 2009; Agarwal and Yu, 2010)) and reference scope identification in citing sentences (Abu-Jbara and Radev, 2011; Abu-Jbara and Radev, 2012).

5 Conclusions and Future Work

In this paper we explore the role of triggers and their scopes in biographical fact extraction. We implement the trigger scoping strategy using two simple but effective methods. Experiments demonstrate that our approach outperforms state-of-the-art without any syntactic analysis and external knowledge bases.

In the future, we will aim to explore how to generate a trigger list for a "surprise" new fact type within limited time.

Acknowledgement

This work was supported by the U.S. DARPA Award No. FA8750-13-2-0045 in the Deep Exploration and Filtering of Text (DEFT) Program, the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), U.S. NSF CAREER Award under Grant IIS-0953149, U.S. AFRL DREAM project, IBM Faculty Award, Google Research Award, Disney Research Award, Bosch Research Award, and RPI faculty start-up grant. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

A. Abu-Jbara and D. Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proc. Association for Computational Linguistics (ACL2011)*. Association for Computational Linguistics.

A. Abu-Jbara and D. Radev. 2012. Reference scope identification in citing sentences. In *Proc. Human*

[3] *A poss^{-1} B* means there is a possession modifier relation (poss) between B and A.

Language Technologies conference - North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2012).

S. Agarwal and H. Yu. 2010. Biomedical negation scope detection with conditional random fields. *Journal of the American medical informatics association*, 17(6):696–701.

J. Aguilar, C. Beller, P. McNamee, and B. Van Durme. 2014. A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. *ACL 2014 Workshop on Events*.

Arnold. 1999. Skimming and scanning. In *Reading and Study Skills Lab*.

C. Chang and C. Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Z. Chen, S. Tamang, A. Lee, X. Li, W. Lin, M. Snover, J. Artiles, M. Passantino, and H. Ji. 2010. Cuny-blender tac-kbp2010 entity linking and slot filling system description. In *Proc. Text Analysis Conference (TAC 2012)*.

D. Douglas and S. Frazier. 2001. Teaching by principles: An interactive approach to language pedagogy (2nd ed.). *TESOL Quarterly*, 35(2):341–342.

P. Elkin, S. Brown, B. Bauer, C. Husser, W. Carruth, L. Bergstrom, and D. Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13.

R. Grishman. 2013. Off to a cold start: New york universitys 2013 knowledge base population systems. In *Proc. Text Analysis Conference (TAC 2013)*.

H. Ji, R. Grishman, H. Dang, K. Griffitt, and J. Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Proc. Text Analysis Conference (TAC 2010)*.

H. Ji, R. Grishman, and H. Dang. 2011. An overview of the tac2011 knowledge base population track. In *Proc. Text Analysis Conference (TAC 2011)*.

Y. Li, S. Chen, Z. Zhou, J. Yin, H. Luo, L. Hong, W. Xu, G. Chen, and J. Guo. 2012. Pris at tac2012 kbp track. In *Proc. of Text Analysis Conference (TAC 2012)*.

Y. Li, Y. Zhang, D. Li, X. Tong, J. Wang, N. Zuo, Y. Wang, W. Xu, G. Chen, and J. Guo. 2013. Pris at tac2013 kbp track. In *Proc. Text Analysis Conference (TAC 2013)*.

C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. Association for Computational Linguistics (ACL2014)*.

B. Min, X. Li, R. Grishman, and A. Sun. 2012. New york university 2012 system for kbp slot filling. *Proc. Text Analysis Conference (TAC 2012)*.

R. Morante and W. Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proc. ACL 2009 Workshop on Current Trends in Biomedical Natural Language Processing*.

K. Motallebzadeh and N. Mamdoohi. 2011. Language learning strategies: A key factor to improvement of toefl candidates reading comprehension ability. *International Journal of Linguistics*, 3(1):E26.

R. Phipps. 1983. *The Successful Student's handbook: A Step-By-Step Guide to Study, Reading, and Thinking Skills.* Seattle and London: University of Washington Press.

B. Roth, T. Barth, M. Wiegand, M. Singh, and D. Klakow. 2013. Effective slot filling based on shallow distant supervision methods. *Proc. Text Analysis Conference (TAC 2013)*.

A. Sun, R. Grishman, W. Xu, and B. Min. 2011. New york university 2011 system for kbp slot filling. In *Proc. Text Analysis Conference (TAC 2011)*.

M. Surdeanu and H. Ji. 2014. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proc. Text Analysis Conference (TAC2014)*.

M. Surdeanu, D. McClosky, J. Tibshirani, J. Bauer, A. Chang, V. Spitkovsky, and C. Manning. 2010. A simple distant supervision approach for the tac-kbp slot filling task. In *Proc. Text Analysis Conference (TAC 2010)*.

M. Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *Proc. Text Analysis Conference (TAC 2013)*.

G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. 2008. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proc. ACL Workshop on Current Trends in Biomedical Natural Language Processing*.

D. Yu, H. Li, T. Cassidy, Q. Li, H. Huang, Z. Chen, H. Ji, Y. Zhang, and D. Roth. 2013. Rpi-blender tac-kbp2013 knowledge base population system. In *Proc. Text Analysis Conference (TAC 2013)*.