Iteratively Estimating Pattern Reliability and Seed Quality With Extraction Consistency [*]

Yi-Hsun Lee[a,b], Chung-Yao Chuang[b], and Wen-Lian Hsu[a,b]

[a]Institue of Information Systems and Applications, National Tsing Hua University,,
101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013, R.O.C.
[b]Institute of Information Science, Academia Sinica,
128 Academia Road, Sec.2, Nankang, Taipei, Taiwan, ROC
{rog, cychuang, hsu}@iis.sinica.edu.tw

Abstract. In this paper, we focus on the task of distilling relation instances from the Web. Most of the approaches for this task were based on provided seed instances or patterns to initiate the process. Thus, the result of the extraction depends largely on the quality of the instances and patterns. For this matter, we propose an iterative mechanism that estimates the reliability of a pattern by the consistency of its extractions, and reevaluate the usefulness of seed instance based on estimated pattern reliability. The resulting system is a semi-supervised method that can take a large quantity of seed instances with diverse quality. To evaluate the effectiveness of our approach, we experimented on 8 types of relationships. The empirical results show that our system performs quite consistency in different relationships while maintaining high precision and recall value.

Keywords: consistency estimation, relation extraction, semi-supervised approach, seed quality

1 Introduction

The rapid growth of the World Wide Web has attracted a lot of research effort on designing methods that automatically extract knowledge or useful information from large, unstructured text. Different from the conventional corpus, the magnitude and noisy natural of the Web has prohibited analytical approaches to be effective. Consequently, most of the systems that took this challenge proceed in a semi-supervised fashion with a human-provided starting point, such as a few instances of the desired extraction(Mann and Yarowsky, 2005; Muslea, 1999; Ravichandran and Hovy, 2002; Pantel and Pennacchiotti, 2006).

In this paper, we focus on the extraction of relation instances. In this scenario, the system needs to be fed with prepared pairs, such as *<Barack Obama, Auguest 4th>*, or initial extraction patterns to bootstrap. Kozareva and Hovy (2010) mentioned that seed selection plays an important role in this kind of semi-supervised approaches. Therefore, how to select high quality seeds in the initial stage is a critical issue. Most researchers select seeds manually to avoid this problem, but the scalability of such manual selection is not promising. Then, these approaches utilize the given instances, called seeds, and generate extraction patterns that has the potential to locate more instances of the desired type in the text. For example, Ravichandran and Hovy (2002) use surface text patterns like `<Person> was born on <Date>` to answer questions about birth dates.

[*] This research was supported in part by the National Science Council under grant NSC 100-2631-S-001 -001, and the research center for Humanities and Social Sciences under grant IIS-50-23.

```
{{Infobox NBA Player
| name = Kobe Bryant
| position = [[Shooting guard]]
| team = Los Angeles Lakers
| salary = 23,034,375
| nationality = [[United States|American]]
| birth_date = {{Birth date and age|1978|8|23}}
| birth_place = [[Philadelphia]], [[Pennsylvania]]
| highschool = [[Lower Merion High School|Lower Merion HS]]}}
```

Figure 1: An example of Wikipedia Infobox

Different from those approaches that heavily depend on the quality of the initial seeds, in this paper, we took an alternative direction that focuses more on the quantity of the seed instances. Such a pursuit is made possible by the advent of rich knowledge sources such as Wikipedia[1] and CIA Fact Book[2]. For example, Wikipedia Infoboxes[3] provide an opportunity to easily gather a vast amount of seed instances because the data is stored in a template form as illustrated in Figure 1. Using such sources, we can harvest a large number of pairs like <*Kobe Bryant, Pennsylvania*> from the below infobox as seed instances for birth place extraction. However, using arbitrary seeds to retrieve sentences from the Web will potentially result in a large number of irrelevant content, which will in turn hamper pattern production.

To demonstrate such a situation, we conducted an experiment on seeds gathered from Wikipedia infobox. The results are presented in Table 1. For each relation type, we randomly select 200 seed instances for forming queries and for each query, evaluate first ten snippets returned from search engine. The relevance of the retrieved snippets are judged by two human annotators. We can see that the relevance ratio is not perfect even we have used both entities in the pair for forming the query. One factor behind such imperfection is that the open and voluntary nature of Wikipedia allows editors to fill in information of different specificity. For example, the birth place field of some people contains only less detailed information such as the country instead of more specific description like county or city. Moreover, as can be seen in Table 1, the relevance ratio is not consistent among different relation types and may be surprisingly low such as the type of death place. Such a situation will affect the performance of semi-supervised approaches greatly(Xu *et al.*, 2007).

Fortunately, having abundant seed instances offers us an opportunity to mitigate such a problem. In this paper, we propose a mechanism that iteratively assesses both the quality of seed instances and induced extraction patterns. Our strategy is to estimate the reliability of an extraction pattern by the consistency of its extractions, and alternately, reevaluate the usefulness of seed instances based on estimated pattern reliability. The resulting system works best when it is fed with a large number of seeds, so that the reliability of the induced pattern can be better estimated.

In the next section, we review several semi-supervised approaches that are comparable

[1] http://www.wikipedia.org

[2] http://www.cia.gov/library/pulications/the-world-factbook

[3] http://en.wikipedia.org/wiki/Help:Infobox

Table 1: Relevance ratio of snippets retrieved by submitting seed instances as queries.

Type	Birth Date	Birth Place	Birth Year	Death Date	Death Place	Death Year	Nobel Prize	Spouse
Ratio	0.96	0.768	0.955	0.939	0.107	0.888	0.879	0.889

to our system. We introduce the proposed CEPRA method in Section 3. In Section 4, we describe the experimental settings; and in Section 5, we discuss the experiments conducted to evaluate the performance of different selection approaches. We summarize the results in Section 6. Then, in Section 7, we provide some concluding remarks and consider avenues for future research.

2 Related Works

Semi-supervised approaches start with manually prepared patterns or seeds, and then generate surface text patterns, which are syntactic patterns that connect two entities in one relationship. Surface text patterns are widely used for information extraction. For example, `<Person> was born in <Year>` is an intuitive pattern for matching the birth year of someone. This pattern connect the person and the corresponding year as a semantic relation (birth year) and thus can be used to effectively extract the information. For binary relation extraction, such as the above example of *<Person, Birth Year>*, the first term is often called the *hook* term(e.g. *Person*), and the second one the *target* term(e.g. *Birth Year*) (Alfonseca *et al.*, 2006; Mann and Yarowsky, 2005; Ravichandran and Hovy, 2002).

Most pattern-based approaches for relation extraction are implemented as follows. First, a set of seed instances are prepared in the form of pairs of hook and target terms serving as examples of the intended relation type. For instance, *<Kobe Bryant, 1978>* could be one of the seed instances that fed into a relation extraction system for learning how to extract the birth year of someone. The seed instances are then used as queries for retrieving sentences containing both the hook and the target terms, most popularly from the Web. The retrieved sentences are subsequently used for generating extraction patterns. Several approaches for this step have been proposed such as the longest common substring (Agichtein *et al.*, 2001), substrings in suffix trees (Ravichandran and Hovy, 2002; Ruiz-Casado *et al.*, 2007) and edit distance based alignment (Ruiz-Casado *et al.*, 2007). However, a portion of the retrieved sentences can be, to a certain degree, not relevant in describing the intended relation. Thus, the patterns built on top of them can deviate from the original goal. To overcome this problem, most approaches place an evaluation step in which patterns are assessed using certain criteria. Such an evaluation is usually done using a sentence collection different from the one used for producing the extraction patterns. When using the Web as our extraction source, one apparent choice of this testing collection is the sentences retrieved by only submitting the hook terms to the search engine. In this paper, we denote such a sentence collection as $\mathbf{S}_H$, and the sentence collection gathered using both hook and target terms as $\mathbf{S}_{\langle H,T\rangle}$.

With an additional sentence collection, we can assess the utility of the generated patterns. A simple way to estimate the usefulness of an extraction pattern, p, is based on the extraction frequency of that pattern. This frequency-based estimation (FE) method counts how many terms (regardless correct or not) that pattern, with hook-term slot filled, is able to extract from $\mathbf{S}_H$,

$$f_{FE}(p) = \sum_{h \in \mathbf{H}} |p_h(\mathbf{S}_H)|$$

where $\mathbf{H}$ is the set of all hook terms, p_h is the pattern p with hook term h, $p_h(\mathbf{S}_H)$ is the bag of terms that p_h extracted from $\mathbf{S}_H$, and $|p_h(\mathbf{S}_H)|$ denotes the size of that bag of terms.

Another intuitive approach for evaluating extraction patterns is based on estimated accuracy. The accuracy can only be estimated because of the noisy nature of the Web. For example, in the Chinese portion of Wikipedia, the birth place of Kobe Bryant extracted from infoboxes is “賓夕法尼亞州 (Pennsylvania)” and “費城 (Philadelphia)”. However, there are several translations other than the above for “Pennsylvania” and “Philadelphia” in Chinese, which all could appear in the retrieval. Besides the translation problem that we encountered frequently in this work, the voluntary nature of Wikipedia renders automatic evaluation vulnerable to specificity mismatches. For instance, the extracted information may be more detailed than the information provided in the infobox, however, there is no general and convenient way to adjust such a mismatch. For these reasons, rather than calling this approach accuracy-based, we refer it as confidence-based estimation (CE), and is formulated as

$$f_{CE}(p) = \frac{\sum_{\langle h,t\rangle \in \langle \mathbf{H},\mathbf{T}\rangle} |p_h(\mathbf{S}_H) = t|}{\sum_{h\in \mathbf{H}} |p_h(\mathbf{S}_H)|}$$

where $\langle \mathbf{H}, \mathbf{T}\rangle$ is the set of seed instances in which each $\langle h, t\rangle$ is a pair of hook and target terms, and $|p_h(\mathbf{S}_H) = t|$ is the number of terms in $p_h(\mathbf{S}_H)$ that matches t. Both frequency and confidence-based approaches rely heavily on the seed quality. In order to alleviate that, Pantel and Pennacchiotti (2006) proposed a method called *Espresso* which uses point-wise mutual information (PMI) to evaluate the strength of association between a pattern and its extractions,

$$f_{ES}(p) = \frac{\sum_{\langle h,t\rangle \in \langle \mathbf{H},\mathbf{T}\rangle} \frac{pmi(\langle h,t\rangle,p)}{\max_{pmi}} \times g_{ES}(\langle h,t\rangle)}{|\langle \mathbf{H},\mathbf{T}\rangle|}$$

where $pmi(\langle h, t\rangle, p)$ is the point-wise mutual information between pattern p and a relation instance $\langle h, t\rangle$, $\max_{pmi}$ is the largest PMI observed, and $g_{ES}(\langle h, t\rangle)$ is an estimate of the quality of instance $\langle h, t\rangle$,

$$g_{ES}(\langle h,t\rangle) = \frac{\sum_{p\in \mathbf{P}} \frac{pmi(\langle h,t\rangle,p)}{\max_{pmi}} \times f_{ES}(p)}{|\mathbf{P}|}$$

where $\mathbf{P}$ is the set of all patterns. These two formulas are calculated iteratively to adjust the weights of both patterns and seed instances. This approach utilizes the co-occurrence as an indication. However, patterns frequently co-occurred with some instances may still have no relevance to the targeted relation. For example, in (Blohm *et al.*, 2007), *Espresso* got a low precision result in birth year extraction. Xu *et al.* (2007) noted that a factor for pattern-based approach to be effective is the ratio of relevant sentences within the text collection for generating the patterns. In this paper, we gather many seeds extracted from Infoboxes in the initial stage. Due to the seed quality is not consistent, we propose a new approach for evaluating both the patterns and seed instances. Different from the above approaches which only evaluate performance based on $\mathbf{S}_H$, our proposal further utilizes the statistical similarity between extractions from $\mathbf{S}_{\langle H,T\rangle}$ and extractions from $\mathbf{S}_H$, which we believe is a good indicator of pattern reliability.

3 Proposed Approach

In this section, we focus on the main concept about estimating pattern's reliability with consistency measurement between different sentence collections and measure the seed's quality by reliable patterns, described in the following sub-sections.

As pattern generation process can potentially produce a large number of extraction patterns, we need a strategy to find the most reliable patterns. To explain the intuition behind our approach, consider an “oracle collection”, $\mathbf{S}_O$, which contains all sentences describing the targeted relation that we can find on the Web. Figure 2 shows the relationship between $\mathbf{S}_H$, $\mathbf{S}_{\langle H,T\rangle}$ and $\mathbf{S}_O$. Ideally, $\mathbf{S}_O$ and $\mathbf{S}_{\langle H,T\rangle}$ would be subsets of $\mathbf{S}_H$ if we could retrieve all sentences containing hook terms from the Web. As depicted in Figure 2, $\mathbf{S}_{\langle H,T\rangle}$

is not a subset of $\mathbf{S}_O$ because the sentences retrieved by using hook and target terms will contain some noisy results as discussed in Section 1 and demonstrated in Table 1. When generating patterns, $\mathbf{S}_{\langle H,T\rangle} \setminus \mathbf{S}_O$ will cause relevance problems, which makes the pattern induction procedure producing deviated patterns. On the other hand, $\mathbf{S}_O \setminus \mathbf{S}_{\langle H,T\rangle}$ will cause specificity problems, which undermines useful patterns in evaluation.

Manually selecting seed instances could reduce the extent of these two kinds of problems. However, the scalability of such an approach is not promising. Thus, we need an approach to automatically assess the utility of each seed instance in evaluating the extraction patterns. Supposedly, high utility seed instances are the ones which distribute high scores into the reliable patterns. In this work, we assume that the reliability of a pattern can be measured by looking into the performance similarities between applying that pattern to $\mathbf{S}_H$ and $\mathbf{S}_{\langle H,T\rangle}$. We consider such a similarity because if a pattern extract mostly in the intersection of $\mathbf{S}_H$, $\mathbf{S}_{\langle H,T\rangle}$ and $\mathbf{S}_O$, then its performance will be consistent regardless extracting from $\mathbf{S}_H$ or $\mathbf{S}_{\langle H,T\rangle}$. Once those highly reliable patterns are found, we could in turn assessing the quality of each seed instance.

In this work, we use the *extended Jaccard coefficient*(Strehl and Ghosh, 2000) (EJAC) to evaluate the consistency in performance when applying to two different sentence collection

$$J(p) = \frac{V_{\mathbf{S}_H}(p) \bullet V_{\mathbf{S}_{\langle H,T\rangle}}(p)}{\parallel V_{\mathbf{S}_H}(p) \parallel^2 + \parallel V_{\mathbf{S}_{\langle H,T\rangle}}(p) \parallel^2 - V_{\mathbf{S}_H}(p) \bullet V_{\mathbf{S}_{\langle H,T\rangle}}(p)} \qquad (1)$$

where $V_X(p)$ is the performance vector of p under sentence collection X which comprises the pattern's weighted precision estimates for each seed instance

$$V_X(p) = \left(\ \lambda_{\langle h_1,t_1\rangle}(p,X), \quad \lambda_{\langle h_2,t_2\rangle}(p,X), \quad ..., \quad \lambda_{\langle h_n,t_n\rangle}(p,X)\ \right)$$

where $\lambda_{\langle h_i,t_i\rangle}(p,X) = A_{\langle h_i,t_i\rangle}(p,X) \times w_{\langle h_i,t_i\rangle}$ in which $w_{\langle h_i,t_i\rangle}$ is the weight of seed instance $\langle h_i, t_i\rangle$ (described below), and $A_{\langle h_i,t_i\rangle}(p,X)$ is the estimated precision of p,

$$A_{\langle h_i,t_i\rangle}(p,X) = \frac{|p_{h_i}(X) = t_i|}{|p_{h_i}(X)|}$$

When binding to some hook terms, a pattern may not be able to extract anything from $\mathbf{S}_H$ or $\mathbf{S}_{\langle H,T\rangle}$. In this case, we use the *expected target accuracy* (ETA), μ_X, as the missing value,

$$\mu_X(p) = \frac{\sum_{\langle h,t\rangle \in \langle \mathbf{H},\mathbf{T}\rangle} A_{\langle h,t\rangle}(p,X) \times w_{\langle h,t\rangle}}{\sum_{\langle h,t\rangle \in \langle \mathbf{H},\mathbf{T}\rangle} w_{\langle h,t\rangle}} \qquad (2)$$

where X can be $\mathbf{S}_H$ or $\mathbf{S}_{\langle H,T\rangle}$. However, using only Equation 1 is not sufficient because inferior patterns would have similar low precision distribution in both $\mathbf{S}_H$ and $\mathbf{S}_{\langle H,T\rangle}$.

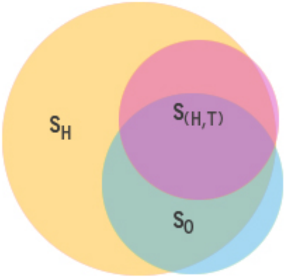


Figure 2: Overlap Diagram between $\mathbf{S}_H$, $\mathbf{S}_{\langle H,T\rangle}$ and $\mathbf{S}_O$

Therefore, we combine Equation 1 and 2 to form our evaluation metric

$$f_{CEPRA}(p) = \alpha \times \mu_{\mathbf{S}_{\langle H,T \rangle}}(p) + \beta \times J(p)$$

In this paper, we set both α and β to 0.5.

As mentioned earlier, the quality of the seed instances is an important aspect of this task. Intuitively, the utility of a seed instance can be assessed by the frequency that it is matched by extraction patterns. This assumes that the more inbound links to a seed, the higher quality it gets. However, there may be some overly-general patterns that are characterized by high coverage and low precision. Therefore, we need to prune those possibly unreliable patterns and utilize the remaining ones to evaluate the quality of seed instances. It is conceivable that if the reliable patterns cannot extract correct target term for a specific hook term, then the quality of this seed instance is questionable. In other word, the more reliable the patterns matching the instance, the higher will be the quality of that instance. Hence, we derive our formula for weighing seed instance as

$$v_{\langle h,t \rangle} = \sum_{p \in \mathbf{P}} \begin{cases} A_{\langle h,t \rangle}(p, \mathbf{S}_{\langle H,T \rangle}), & \text{if } f_{CEPRA}(p) > \varepsilon \\ 0, & \text{otherwise} \end{cases}$$

where $\varepsilon = 0.7$ in this work. This value is normalized to obtain the final weight

$$w_{\langle h,t \rangle} = \frac{v_{\langle h,t \rangle} - \min_i v_{\langle h_i,t_i \rangle}}{\max_i v_{\langle h_i,t_i \rangle}/c}$$

where $c = 20$ in this work.

$w_{\langle h,t \rangle}$'s and $f_{CEPRA}(p)$'s are calculated iteratively. Initially, we compute $f_{CEPRA}(p)$'s using $w_{\langle h,t \rangle} = 1$. Then, we collect patterns with high f_{CEPRA} values to determine $w_{\langle h,t \rangle}$'s. After setting the seed weights, we can re-calculate the f_{CEPRA}'s with reduced influence from inferior seeds. As this process iterates, only patterns whose f_{CEPRA} values higher than a threshold are retained. Note that we also keep patterns with high ETA value if it can extract several seed instances with high $w_{\langle h,t \rangle}$ and their f_{CEPRA} is set to its $\mu_{\mathbf{S}_{\langle H,T \rangle}}$.

4 System Architecture and Experimental Environment

In this section, we describe our system, Consistent Estimation Pattern-based Relation Acquirer (CEPRA), which runs briefly as follows. Based on the seed instances gathered from the infoboxes in Chinese portion of Wikipedia, we collect two sentence collections $\mathbf{S}_H$ and $\mathbf{S}_{\langle H,T \rangle}$ from the Web. The retrieved sentences are preprocessed with segmentation and parts-of-speech tagging. Next, extraction patterns are generated with an alignment-based approach (Sung *et al.*, 2009) based on sentences in $\mathbf{S}_{\langle H,T \rangle}$. Those patterns are then evaluated by the approach described in Section 3. Finally, we utilize the retained patterns to extract relation instances.

To assess the performance of our method, we performed experiments on 8 types of relations. The experimental settings are described below:

Experiment Data: We collected our seed instances from the infoboxes in the Chinese portion of Wikipedia. The collected instances are biographical relations and of 8 different types. To compile the training $\mathbf{S}_{\langle H,T \rangle}$, for each relation type, we used 1000 seed instances and submitted the hook and target terms of each training seed to retrieve 50 snippets from Google. Among those 1000 seed instances, we randomly picked 50 instances and used their corresponding sentences retrieved from the Web to run pattern generation. The produced patterns are evaluated using whole training sentence collection. A testing sentence collection[4] is formed by using a separate set of 200 instances from each relation type. This testing set contains 3768 sentences and is annotated by two human annotators (the relevance statistics are shown in Table 1.)

Evaluation and Comparisons: We compared our method with three other approaches described in Section 2. In this paper, we want to evaluate the difference in pattern selection, thus, we use the same pattern generation procedure for all 4 approaches. The results are compared both in recall and precision.

5 Evaluation of Pattern Estimation Approaches

Figure 3 shows the precision value of the eight relationships for top N ranked patterns. In this figure, we can observe that none of the approaches performs well on the *Death Place* relationship. In Table 1, we can find although the relevant ratio is quite low in the *Death Place* relationship, which affects the performance of the pattern-based systems. However, on this relationship, our approach achieves a higher precision score (0.67) than the compared approaches. In the *Birth Place* relationship, the FE, CE and *Espresso* approaches would get a lower precision at top 500 ranked patterns. In contrast, our approach achieves a perfect precision score of 1 on the top 500 and 1000 patterns. In these relationships with lower relevant ratio, our approach still performs better than other approaches. Next, in some relationships with high quality seeds like *Birth Date, Birth Year, Nobel Prize* and *Spouse*, the CE and our approach both achieve good performance. But we can find the CE approach with a significant drop when N increased in *Death Date* and *Death Year* relationships. Contrast to our approach, we still get a stable and good performance, higher than 0.9, in these relationships. It means that our approach would gather more reliable patterns than other approaches. Finally, unlike other approaches, the *Espresso* approach only considers the relationships between the targets and the patterns. It did not perform well in our experiments.

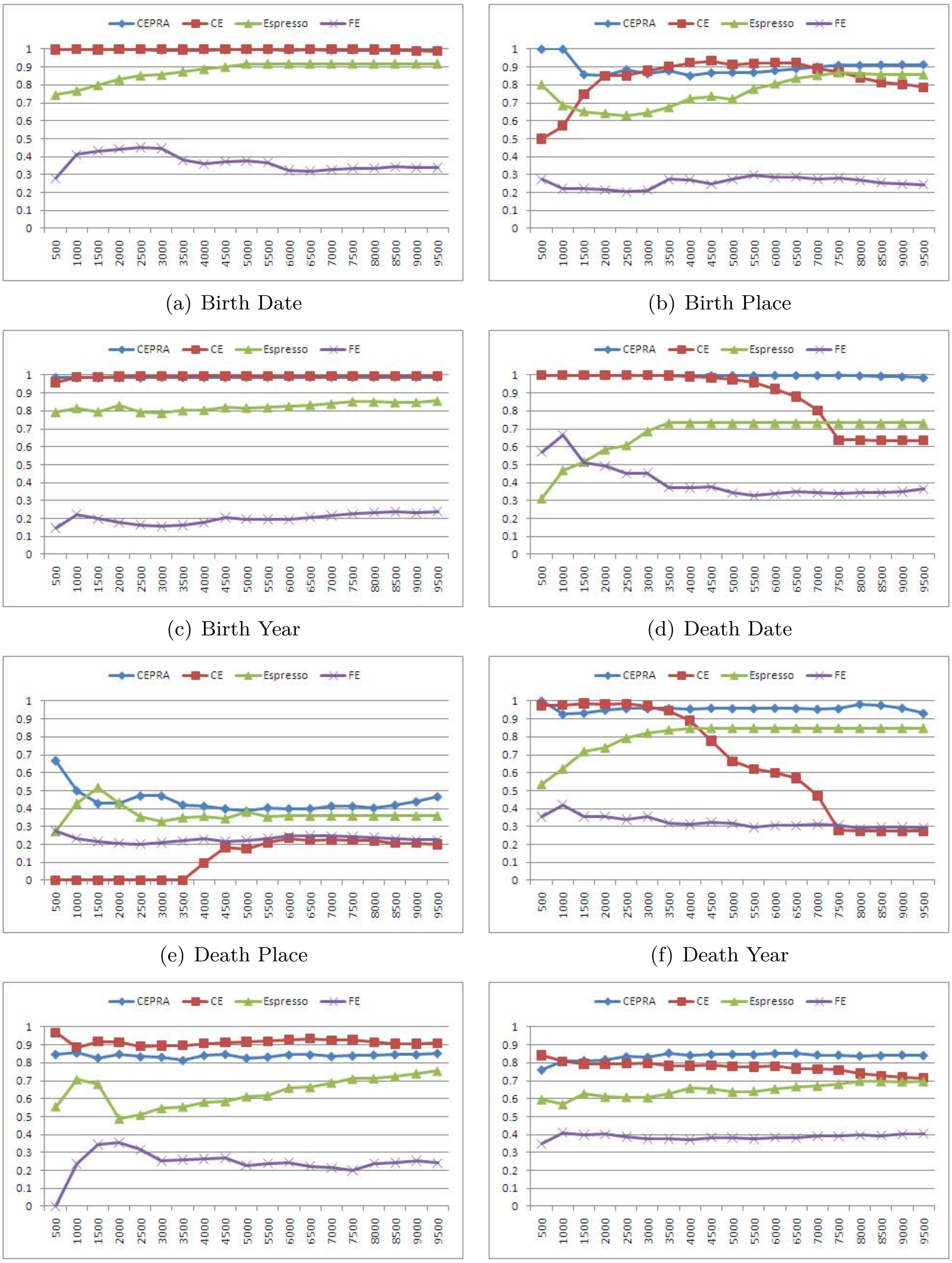
Next we observe the methods' distribution of precision value between these eight relationships is quite different In Figure 3. Our approach achieves the highest precision score on the top 500 ranked patterns, but the score decreases slightly as N increases. In contrast, the *Espresso*'*s* precision score increases with N ranked patterns. However, this phenomenon presents that the top N ranked patterns would not be the good choice. Figure 3 also shows that the CE approach performs quite unstable in different relationships. This approach is easily influenced on the quality of seeds. Next, we discuss the recall and precision value derived by the compared approaches on the eight relationships.

Table 2 shows the highest recall value of the eight relationships with precision value above γ, range from 0.7 to 0.9. Because the *Espresso* and FE approaches get lower

[4] http://[removed for review]/cepra/

Table 2: Highest recall value with different precision threshold, γ, in these eight relationships.

Type	Method	γ			Type	Method	γ		
		0.9	**0.8**	**0.7**			**0.9**	**0.8**	**0.7**
Birth Date	CE	0.68	0.70	0.74	Birth Place	CE	0.17	0.35	0.43
	CEPRA	0.71	0.74	0.74		CEPRA	0.17	0.35	0.43
Birth Year	CE	0.71	0.73	0.73	Death Date	CE	0.57	0.58	0.58
	CEPRA	0.71	0.72	0.73		CEPRA	0.56	0.58	0.58
Death Place	CE	0	0	0	Death Year	CE	0.28	0.33	0.35
	CEPRA	0	0	0		CEPRA	0.33	0.41	0.43
Nobel Prize	CE	0.39	0.42	0.43	Spouse	CE	0	0.14	0.43
	CEPRA	0.38	0.42	0.43		CEPRA	0	0.4	0.48



(a) Birth Date

(b) Birth Place

(c) Birth Year

(d) Death Date

(e) Death Place

(f) Death Year

(g) Nobel Prize

(h) Spouse

Figure 3: Precision of the Top N ranking patterns of the four compared approaches, *CEPRA*, *CE*, *FE* and *Espresso*. The x-axis represents the precision value and the y-axis represents the number of N.

precision values at top N ranked patterns, we do not consider these two approaches in this table. First of all, in the *Death Year* relationship, we can find our approach achieves a higher recall value than the CE approach. As shown in Figure 3, the CE would encounter a significant drop in this relationship at top 6000 to 7000 ranked patterns. Compared to our approach, we not only retain more reliable patterns but we also get a higher recall value. In the *Spouse* relationship, although our approach get a bit lower precision value, we get a higher recall value 0.4 compared to 0.14 for the CE approach. Generally speaking, in this empirical experiment, our approach would retain more reliable patterns while maintaining stable and high performance in different relationships.

6 Discussion

As mentioned in Sections 5, the performance of CE approach is not stable on different relationships. We assume that the problem is caused by 1) insufficient information to judge whether the target is correct, which means that the CE approach yields a biased confidence score because of the influence of low quality seeds; 2) the lack of relevant sentences in the $\mathbf{S}_H$. In the FE approach, the precision value of the top N patterns is quite low because a pattern with a high frequency is always a general pattern. In some domains, highly frequent patterns may be useful for finding new instances. However, in relation extraction tasks, we need to find some related or specific patterns instead of highly frequent patterns. Next, we consider the performance of the *Espresso* approach. In this paper, we utilize collections comprised of sentences retrieved from the Web. This factor may affect the performance of the *Espresso* approach. However, the patterns highly occurred with some targets possible describe different relationships. Besides, how to compile a training set is an important issue when utilizing the *Espresso* approach. In this paper, we propose an approach to estimate a pattern's reliability between different sets. We use an iteratively pattern-seed evaluation method to prune irrelevant patterns and low quality seeds. In Sections 5, we noted that the $CEPRA$ achieved a stable performance on different relationships. However, we still need to resolve the following issues. 1) Because our approach was based on utilizing many seeds extracted from Wikipedia Infoboxes, we need to observe more details about the seed sets like the learning curve with different sizes of seed sets or the performance on different ratio of inferior quality seeds. 2) Currently, we do not consider the issue about pattern generation. Maybe we need to concern about utilizing more complex linguistic tools to generate frequent patterns and check our validation performance is consistent or not.

7 Conclusion

Unlike other semi-supervised approach utilizing manually prepared seeds in the beginning, we proposed a method to estimate pattern's reliability by abundant seeds extracted from Wikipedia Infoboxes automatically. First, we employ an automatic approach to select sentences, and then use an alignment-based pattern generation approach. Next we apply consistency measurement to estimate pattern's reliability and utilize an iteratively approach to find high quality seeds and reliable patterns. Finally, we use the derived patterns to find precise targets. Based on our experimental result, our system has a more stable and better performance than other approaches. In our future work, we will discuss more experiments about the sizes of seed sets and how to utilize a deep linguistic tool to improve the system's performance.

References

Agichtein, Eugene, Luis Gravano, Jeff Pavel, Viktoriya Sokolova, and Aleksandr Voskoboynik. 2001. Snowball: a prototype system for extracting relations from large text collections. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, SIGMOD '01, p. 612, New York, NY, USA. ACM.

Alfonseca, Enrique, Pablo Castells, Manabu Okumura, and Maria Ruiz-Casado. 2006. A rote extractor with edit distance-based generalisation and multi-corpora precision calculation. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pp. 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Blohm, Sebastian, Philipp Cimiano, and Egon Stemle. 2007. Harvesting relations from the web: quantifiying the impact of filtering functions. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, pp. 1316–1321. AAAI Press.

Kozareva, Zornitsa and Eduard Hovy. 2010. Not all seeds are equal: measuring the quality of text mining seeds. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 618–626, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mann, Gideon S. and David Yarowsky. 2005. Multi-field information extraction and cross-document fusion. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 483–490, Stroudsburg, PA, USA. Association for Computational Linguistics.

Muslea, Ion. 1999. Extraction patterns for information extraction tasks: A survey. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, pp. 1–6.

Pantel, Patrick and Marco Pennacchiotti. 2006. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pp. 113–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ravichandran, Deepak and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 41–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ruiz-Casado, Maria, Enrique Alfonseca, and Pablo Castells. 2007. Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data Knowl. Eng.*, 61, 484–499, June.

Strehl, Alexander and Joydeep Ghosh. 2000. Value-based customer grouping from large retail data-sets. In *In Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery*, pp. 33–42.

Sung, Cheng-Lung, Cheng-Wei Lee, Hsu-Chun Yen, and Wen-Lian Hsu. 2009. Alignment-based surface patterns for factoid question answering systems. *Integr. Comput.-Aided Eng.*, 16, 259–269, August.

Xu, Feiyu, Hans Uszkoreit, and Hong Li. 2007. A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 584–591, Prague, Czech Republic, June. Association for Computational Linguistics.