Extracting Predictive Statements with Their Scope from News Articles

Navya Yarrabelly
Data Sciences and Analytics Center
IIIT-Hyderabad
yarrabelly.navya@research.iiit.ac.in

Kamalakar Karlapalem
Data Sciences and Analytics Center
IIIT-Hyderabad
kamal@iiit.ac.in

Abstract

We estimate that a large number of news articles contain references to future. The reference is detected through the notion of predictive statements (phrases). Distinguishing such predictive statements from factual statements in news articles is important for most applications such as fact checking, opinion mining, future trend analysis, etc. In this paper, we approach the problem of automatically extracting future-related information by solving two sub-problems. The first sub-problem is labeling a sentence as predictive or factual. In addition to extracting the predictions, we address the tasks of clausal scope resolution and dis-embedding linguistic peripheral clauses with respect to the predictive clause in a sentence. To solve these problems, we extract all the clauses of a given sentence and classify each of the clauses as predictive or factual. We then use a machine learning based approach to disambiguate the clause labels by using the clausal dependency relations and label the sentence.

1 Introduction

We define prediction[1] as any statement made in reference to future i.e, any statement about what will or might happen in the future. In newspaper articles, many journalists evaluate the current state of affairs and predict possible future scenarios. In some sense, the journalist is rated high based on their ability to evaluate and predict the future scenarios with some degree of assurance. Therefore, it is imperative to determine the passages, sentences and phrases of the news articles that predict the future scenarios. A person well versed with reading articles can easily determine predictability aspects of a news article and over time, has some assurance about which articles or news agencies correctly predict some of the future scenarios. The key issue is that predictions have a complex interplay between the prediction component of a sentence, fact, and the truthfulness of a fact. An information retrieval and extraction system must handle the above-mentioned interplay in a systematic manner to address the problem of determining predictive statements and their scope in news articles. In our context, we define factual base for a prediction as knowledge, facts, science, experiments etc, based on which the prediction is being made.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]https://www.merriam-webster.com/dictionary/prediction

A sentence with a predictive clause can be predictive or factual depending on its association with other clauses in the sentence. For example, consider three excerpts below, extracted from BBC [2] news stories and Times Now[3] news stories.

I Government promised to extend the maternity leave by 2 months by the end of 2005.

II Though the government promised to extend the maternity leave, it could not keep its promise.

III Rajnath Singh told he believed that, with key pre-poll pacts now in place around the country, the party and its allies could win 300 seats of the 543 being contested.

Statement I is a prediction while statement II is a fact, though both the statements have a common predictive clause *"Government promised to extend the maternity leave"*. Statement III is a prediction with the predictive clause *"the party and its allies could win 300 seats of the 543 being contested"* and factual base for the prediction *"with key pre-poll pacts now in place around the country"*. From these examples, it is clear that the problem of prediction extraction cannot be solved just by extracting the linguistic patterns or keywords which are predictive. But, there is an imperative need to address the solution with a new NLP perspective for processing of a sentence with its component clauses and analyze the dependencies between the clauses, which other methods lack in their approach.

Dataset Preparation: We selected many representative sentences from a set of English news articles scraped from web(online news sites) and labeled the sentences as predictive or factual. For predictive sentences, predictive pattern of the sentence is identified. For example, the sentence *"Saulius Mikoliunas* ***could also face*** *action after three fans were arrested for throwing coins on the pitch"* is labeled predictive with *'could also face'* as the predictive phrase. From such sentences, we created two datasets: Set3480 from Politics and Economy domains and Set200 from Sports domain. Set3480 has 1740 predictive clauses from 1227 predictive sentences, to serve as positive instances and 1740 factual clauses from 1011 factual sentence, to serve as negative instances for the classifier. Set200 has 100 predictive clauses

[2]http://www.bbc.com/news

[3]http://www.timesnow.tv/

from 70 predictive sentences and 100 factual clauses from 58 factual sentences.

2 Related Work

To the best of our knowledge, although linguistically expressed references to the future has been studied by a number of researchers, the problem of extracting the scope of a prediction from a sentence is not addressed by any of the existing work.

(Jatowt et al. 2009; Kawai et al. 2010) extracts and retrieves time-referenced predictions from a given arbitrary query. (Kanhabua, Blanco, and Matthews 2011) retrieves and ranks predictions that are relevant to a news article. These methods suffer from low recall, as we estimate from our results that only 35% of the predictions are time referenced.

CONLL 2010 shared tasks(Farkas et al. 2010)aimed at identifying sentences in texts which contain unreliable or uncertain information. And CONLL 2010 shared task-2 aimed at the resolution of in-sentence scopes of hedge cues in Scientific Text. Our problem distinctly differs from these tasks, as we classify sentences as predictive if it refers to an action or probability of a future course of an event, while the CONLL task classifies sentences to assess the factual degree of events in text.

(Özgür and Radev 2009) detects speculations and their scopes in Scientific Text. It classifies the potential keywords as real speculation keywords or not by using a diverse set of linguistic features that represent the contexts of the keywords. As shown in above example, in the case of news articles, it is syntactic patterns of a sentence coupled with its structure that makes it predictive, more than the mere presence of keywords. (Nakajima et al. 2014; 2015) generates a list of morpho-semantic patterns that appear uniquely in only future-related information. However, their methods suffered from low precision as it considered a whole sentence as a prediction to classify and extract future-specific patterns. As mentioned from above examples(I, II and III), a sentence can contain both predictive and factual parts. Hence factual and predictive parts should be separated and processed individually to extract future specific patterns more accurately.

Unlike previous studies which treat the problem of extracting predictions as sentences which contain expressions or patterns referring to future, we extract features by considering only the predictive parts of a statement while training the model. We also deal with a more complicated problem of identifying the scope, factual base of the prediction from complex-compound sentences.

3 Predictive statement classification

Our framework to classify statements as future predictive, has three components (1) Clausal dependency relations extraction: to split the sentence into its constituent clauses and extract the dependencies of how a clause modifies other clauses in the sentence. (2) Classification of simple clauses(CSC): to label each free-standing clause as predictive or factual, by learning lexical and syntactic patterns of relation triplets in a clause (3) Clause labels disambiguation : a classifier(CCDR) to predict the label of a sentence as predictive or factual, from its clause labels and dependency relations.

3.1 Clausal Dependency Relations

We pre-processed the labeled data and trained a classifier which labels a given clause as predictive or not. For all the pre-processing steps, we have used Stanford Core-NLP module. (Manning et al. 2014)

Clauses Extraction: Using Stanford parse tree, each sentence is split into its constituent dependent and independent clauses. Let $T=\{t_1, t_2 ... t_n\}$ be the set of nodes in the parse tree, corresponding to the words in the predictive phrase(annotated in the dataset). In the parse tree, we find a clausal node say t_i, which has both NP and VP subtrees and is the deepest common ancestor to all the nodes in T. Clauses formed from this node are taken as predictive clauses and other clauses are labeled as factual clauses.

Dependency Relations: Each dependent clause is associated with an independent clause, which it modifies or serves as a component of it. Conjunctions between the clauses signify the relationship between their ideas. We used Penn Discourse TreeBank to classify these connectives to various classes[4] , implying how one clause modifies semantics of other clauses. Predictions spanning over multiple sentences are taken into account, by classifying implicit inter-sentence connectives. We then represent a set of clausal dependency relations in a sentence as follows. *T(P,Q,C)* denotes a relation of type *T* between clause *P* and clause *Q* connected by a conjunction of class *C*. If clause *Q* is dependent on *P*, type *T* denotes dependent clause and if *Q* and *P* are independent clauses, type *T* denotes independent clause. We exploit these clausal dependencies to classify a sentence as predictive or not from the labels assigned to its constituent clauses.

3.2 Classification of Simple Clauses (CSC)

With the intuition that the predictive nature of a sentence is defined by the linguistic patterns contained by its predicates, we extracted features from the predicates in relation triplets to identify the patterns uniquely referring to future.

Triplets Extraction: Each of the clauses extracted above in 3.1, is semantically represented as (subject, predicate, object) triplets, using dependency parse tree of the clause. A clause is represented by more than one relation triplets, if one of the verbs or nouns in the clause introduces a clausal complement (identified by a dependency governed by one of the relations ccomp, xcomp, advcl, acl in the Stanford dependency tree). For example, triplets extracted from the clause *"Sue promised George to respond to his offer"* are t_1= (Sue, **promised**, George) and t_2=(Sue, promised to **respond** to, his offer)

Predictive and Factual Patterns Extraction : In each triplet, the subject and object parts are simply replaced by "subject" and "object". From the processed labeled data, following features are extracted for all the triplets in each clause.

I POS tags: Set of n-grams of POS tags in triplets.

[4]https://github.com/WING-NUS/pdtb-parser/blob/master/sense_levels.png

II Word co-occurrences: Set of n-grams of words in the triplets.

III FTR: Presence of future temporal references in the clause.

IV Events: Presence of explicit and implicit future event references in the clause, annotated using Tarsql toolkit[5], Stanford SUTime [6] and DBPedia entity linking.

V Keyword Dictionary: We manually collected a set of 42 commonly used keywords to refer future event probability.

Let P be the set of features extracted from predictive clauses and F be the set of features extracted from factual clauses. Patterns which occurred with a frequency less than a threshold of 15 are removed from both the sets. Each clause is represented as a feature vector, with features from both P and F. Taking these feature vectors as training set, we trained a classifier to label a clause as predictive or factual.

3.3 Clause Labels Disambiguation A sentence may get both predictive and factual labels from its constituent free-standing clauses. To label a sentence from its constituent clause labels, we modeled a classifier to further label each clause within its context, taking input as its clausal dependency relations extracted(in section 3.1) and clause label when taken without any context(in section 3.2).

Clausal Classification with Dependency Relations (CCDR) : For every clausal dependency relation T(P,Q,C) extracted in section 3, we add an instance represented by a feature vector with **feature set** F: {Predictive clause label for P, Predictive clause label for Q, Clause type of P, Clause type of Q, Clausal Dependency type T, Clausal Connective C between P and Q, Class of C, Dependency Relation labels(from Stanford Enhanced Dependency Graph) governed by predicate phrase of triplets in P with dependents in predicate phrase of triplets in Q}. We have modeled an SVM classifier with these instances to resolve predictive label of a clause by learning its semantic modification with respect to its clausal dependency relations. We use Algorithm 1 to further label a sentence from its clause labels and its dependency relations.

Predictive part, Condition and Base for a Prediction : If a sentence is labeled as predictive and has a clausal dependency relation *T(P,Q,C)*. L_p and L_q be the class labels for the clauses P and Q respectively, then we extract the presence of factual base or condition for the prediction as follows.

1. If L_q is factual and L_p is predictive and class C is Reason, then the predictive part is P and the factual base for the prediction is Q.
2. If L_p is predictive and L_q is predictive and class C is Reason, then the predictive part is P and the condition for the prediction is Q.

Example : Prediction *"Martina Hingis has admitted that Martina Hingis may consider a competitive return to tennis if appearance in Thailand later this month goes well"* has predictive part *"Martina Hingis may consider a competitive return to tennis"*, has condition *"if an appearance in Thailand later this month goes well"* and base *"Martina Hingis has admitted"*.

[5]http://www.timeml.org/index.html

[6]http://nlp.stanford.edu/software/sutime.shtml

Algorithm 1 Predictive Sentence Classification

```
procedure PREDICTIVE SENTENCE CLASSIFIER
Input:
    SC ← Clause labels using CSC
    CDR ← Clausal Dependency Relations
Output:
    S_l ← predictive label for sentence S

    for each clause P in S do
        cp ← predictive label for clause P      ▷ clause P label obtained from classifier CSC
        for each clausal dependency relation T(P,Q,C) of S in CDR do
            F ← feature Vector using feature set F
            cpr ← predictive label for P, modified by Q   ▷ P is labeled w.r.t Q using classifier CCDR
            if cpr is factual then
                cp ← factual
        if cp is predictive then
            S_l ← predictive
```

Feature Set	Precision	Recall	Fscore
Bigrams, FR, Keywords	**0.88**	**0.87**	**0.87**
Trigrams, FR	0.84	0.82	0.83
Bigrams, Trigrams, FR, Keywords	0.87	0.86	0.86
FR	0.54	0.5	0.40

Table 1: Accuracy measures for predictive clauses classification

4 Experiments and Results

In this section, we discuss the experiments performed on the datasets described in Section 1 and the results achieved to verify whether the predictive statement extraction method is effective.

Classification Results : We built linear SVM models using 10 fold cross validation on Set3480, for classification of predictive clauses. Table 1 summarizes the results obtained for the dataset Set3480 for various feature sets(FR refers to future temporal and event references). Table 2 summarizes the results for classification of sentences from Set3480 and Set200, taking input as clause labels and clausal dependencies relations.

As Set200 is extracted from Sports domain and the model is trained on Politics and Economy domains, a high fscore for this dataset shows that our approach to extract predictions works efficiently for any domain. We also tried different classifiers like Bayesian Logistic Regression classifier

Dataset	Precision	Recall	Fscore
Set 200	0.88	0.85	0.86
Set 3480	0.90	0.86	0.87

Table 2: Accuracy for predictive sentences classification

Representation	Precision	Recall	Fscore
Clauses	**0.90**	**0.86**	**0.87**
Sentences	0.89	0.54	0.67
Triplets	0.83	0.62	0.72

Table 3: Accuracy for different representations of predictive part of a prediction.

and also tried with 33% split as test data. Both the classifiers gave roughly the same results (**fscore** around **0.85-0.89**).

Patterns are extracted at different granularity representations of the sentence, to extract predictive features. Table 3 summarizes the classification results for various cases. Sentences in Table 3 refers to the case where a whole sentence is taken as a prediction, instead of taking only the predictive clause of the prediction. Triplets in Table 3 refers to the case, where triplets of a sentence are given as instances to the classifier. Low fscores in these cases imply that extracting clauses to precisely extract the predictive part of a sentence is efficient. From this, it is clear that pre-processing the predictions to extract predictive clauses and extracting linguistic patterns from the triplets, increases efficiency of the prediction extraction model.

4.1 Results for resolving the scope of a prediction

We also annotate a prediction with the predictive part, factual base for the prediction on which it is made and the condition for the validity of the prediction. We randomly selected 181 sentences from Set3480 to manually validate the accuracy of our method. Table 4 summarizes the results on this subset.

Feature	Number of predictions annotated with the feature	Accuracy
Resolving predictive scope	116/181	0.94
Extracting factual base	55/181	0.59
Extracting condition on which prediction depends	24/181	0.74

Table 4: Results for annotating a prediction with predictive scope, factual base, conditional prediction

5 Conclusion and Future Work

We presented an approach to classify a sentence as predictive or factual, by extracting clauses from the sentence and exploiting the dependency structure of the clauses. We learnt linguistic patterns which uniquely refer to future. We tested our method on two datasets of different sizes. We found out that the method performs well for both sets (Fscore around 0.85-0.87). Using these prediction attributes, we further validated the correctness of these predictions, to give credibility scores for authors in the work (Yarrabelly and Karlapalem 2018).

References

Farkas, R.; Vincze, V.; Móra, G.; Csirik, J.; and Szarvas, G. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, 1–12. Association for Computational Linguistics.

Jatowt, A.; Kanazawa, K.; Oyama, S.; and Tanaka, K. 2009. Supporting analysis of future-related information in news archives and the web. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 115–124. ACM.

Kanhabua, N.; Blanco, R.; and Matthews, M. 2011. Ranking related news predictions. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 755–764. ACM.

Kawai, H.; Jatowt, A.; Tanaka, K.; Kunieda, K.; and Yamada, K. 2010. Chronoseeker: Search engine for future and past events. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, 25. ACM.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, 55–60.

Nakajima, Y.; Ptaszynski, M.; Honma, H.; and Masui, F. 2014. Investigation of future reference expressions in trend information. In *Proceedings of the 2014 AAAI Spring Symposium Series, Big data becomes personal: knowledge into meaning–For better health, wellness and well-being*, 31–38.

Nakajima, Y.; Masui, F.; Yamada, H.; Ptaszynski, M.; and Honma, H. 2015. Automatic extraction of references to future events from news articles using semantic and morphological information. In *Proceedings of the 24th International Conference on Artificial Intelligence*, 4385–4386. AAAI Press.

Özgür, A., and Radev, D. R. 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, 1398–1407. Association for Computational Linguistics.

Prasad, R.; Miltsakaki, E.; Dinesh, N.; Lee, A.; Joshi, A.; Robaldo, L.; and Webber, B. L. 2007. The penn discourse treebank 2.0 annotation manual.

Yarrabelly, N., and Karlapalem, K. 2018. Estimating credibility of news authors from their WIKI validated predictions. In *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018.*, 12–17.