Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 137 (2018) 223–234

Procedia Computer Science

www.elsevier.com/locate/procedia

SEMANTiCS 2018 – 14th International Conference on Semantic Systems

Temporal Role Annotation for Named Entities

Maria Koutraki[a,b,*], Farshad Bakhshandegan-Moghaddam[a,b], Harald Sack[a,b]

[a] *FIZ Karlsruhe, Leibniz Institute for Information Infrastructure, Karlsruhe, Germany*
[b] *AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*

Abstract

Natural language understanding tasks are key to extracting structured and semantic information from text. One of the most challenging problems in natural language is ambiguity and resolving such ambiguity based on context including temporal information. This paper, focuses on the task of extracting *temporal roles* from text, e.g. *CEO* of an *organization* or *head* of a *state*. A temporal role has a domain, which may resolve to different entities depending on the context and especially on temporal information, e.g. *CEO* of *Microsoft* in 2000. We focus on the temporal role extraction, as a precursor for temporal role disambiguation. We propose a structured prediction approach based on Conditional Random Fields (CRF) to annotate temporal roles in text and rely on a rich feature set, which extracts syntactic and semantic information from text.

We perform an extensive evaluation of our approach based on two datasets. In the first dataset, we extract nearly 400k instances from Wikipedia through distant supervision, whereas in the second dataset, a manually curated ground-truth consisting of 200 instances is extracted from a sample of *The New York Times* (NYT) articles. Last, the proposed approach is compared against baselines where significant improvements are shown for both datasets.

Keywords: Temporal Role Annotation; Sequence Classification

1. Introduction

A significant part of Web data is represented as text, and thereby, it constitutes one of the most promising resources for further pushing the advances in the field of Semantic Web, especially in extracting structured data. Some of the most notable examples are the knowledge bases (KB) DBpedia [1] and YAGO [19]. Both use information extraction (IE) techniques by harnessing information from Wikipedia's infoboxes and consequentially provide that information in form of RDF triples. Other examples include TextRunner [20], which uses natural language processing (NLP) and IE techniques to extract facts in the form of triples from generic Web documents. This has resulted in many successful approaches towards mapping OpenIE facts to reference KBs like DBpedia [7, 4].

* Corresponding author. Tel.: +49 7247 808 196.
E-mail address: firstname.lastname@fiz-karlsruhe.de

1877-0509
10.1016/j.procs.2018.09.021

Key aspect in all these cases is to map the extracted triples (the subject and object) to named entities from a reference dataset, e.g. Wikipedia. This task is commonly referred to as Named Entity Disambiguation (NED) with state-of-the-art approaches like AIDA [10], or entity linking (EL) like DBpedia Spotlight [15], which try to resolve surface text forms to entities without any type restriction.

While, in case of NED, the focus is only on named entities, which belong to types like `Person`, `Location`, or `Organization`, for EL, there is no restriction of the entity type. However, the models cannot distinguish between surface forms where depending on the temporal context they may refer to different entities. EL models are trained on snapshots of Wikipedia, and as such the model's priors are heavily biased towards resolving surface forms to the most likely entity given the state in a Wikipedia snapshot. Therefore, one main challenge of existing NED/EL methods is that they do not resolve the case where a surface form refers to a *role*, and in particular if such a role also includes a temporal dimension (e.g. *head* of a *state*).

In this paper, we address this open challenge, where we annotate surface text forms in natural language text that represent temporal roles, e.g. *CEO of Microsoft*, *U.S. President*. In general, we denote a role as a specific *position* a person, a location, an organization, or an event can obtain. A role is ambiguous as it may refer to different named entities. To resolve such ambiguity the context and temporal information play a crucial role. For example, to uniquely identify the named entity referred to by *the Pope*, temporal context is required (cf. Fig. 1). Detecting roles in text is a challenging task, as apart from the role ambiguities that may refer to different named entities, the surface text form of a role may overlap with the actual surface form of a named entity, e.g. "*Alexander Pope*"[1] and "*The Pope*", and thereby increase the level of ambiguity.

We propose a temporal role annotation approach based on a structured prediction model via Conditional Random Fields (CRF) [13]. In our model, we perform a sequence classification that aims at annotating words or phrases with the label "*role*" or "*no role*". We rely on syntactic and semantic features that extract local information (e.g. part of speech tag of a word, the word itself etc.) for a word as well as contextual information for words in a window of +/- 2 words (e.g. presence of NE). In addition, we construct specific *local grammar* rules, e.g. "*President of* `LOCATION`", and additionally consider external features extracted from existing knowledge bases. It is worth noting that temporal role annotation is a precursor towards the subsequent role disambiguation step, which resolves the temporal role surface forms to specific named entities. The latest will be subject of future work.

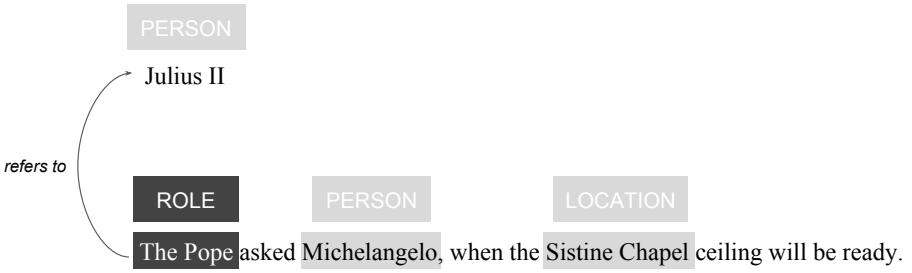

Fig. 1: A role refers to a named entity, but is not identical with a named entity.

To the best of our knowledge, this is the first work that aims at tackling the problem of *temporal role annotation*. Therefore, we construct a large scale ground-truth dataset based on Wikipedia, comprising c. 400k instances of varying difficulty levels, i.e. a difficult instance is considered to be a role, whose surface form may also refer to named entities as e.g. "*Alexander Pope*". Furthermore, to assess how robust our model is across text genres, we show that our pre-trained model on the Wikipedia ground-truth can annotate roles with high accuracy on a small sample of NYT articles.

To this end, the main contributions of this paper are the following:

- the definition of the task of temporal role annotation;

[1] https://en.wikipedia.org/wiki/Alexander_Pope

- the proposal of a structured prediction model for the annotation task; and
- the construction of two high quality ground-truth datasets for this task.

2. Related Work

This section reviews related to the problem of temporal role annotation works.

Semantic Role Labeling (SRL). The work presented in this paper is loosely related to SRL. In SRL the task is to label the arguments of a predicate, usually represented through a verb. The labels can vary in their granularity and can have thematic representations [2], e.g. *AGENT*, *SPEAKER*. Thereby, *"role"* in this sense means a specific semantic class, to which a word or phrase belongs. SRL works like [8] do not apply for our task as they assume the roles to be present only in sentences, where two nouns or noun phrases (commonly referred to as *Protagonists* in SRL) are related through a predicate. In our case, we do not enforce such a restriction since a temporal role does not necessarily have to be part of a relation in a statement.

Named Entity Recognition (NER). In NER, the task is to identify surface forms that resolve to specific named entity categories. State-of-the-art approaches like [6] annotate surface forms with the predefined named entity classes: `Person`, `Organization`, `Place`. While both the NER annotation task and our temporal annotation task share commonalities, in that they annotate words or phrases in a statement with a predefined set of labels, a comparison between NER state-of-the-art methods against our approach is deemed irrelevant. First, NER approaches are optimized to label surface forms that link to the aforementioned named entity classes. In the case of the English language, a common feature is that, if a word begins with a capital letter – a common feature of proper nouns – this is a significant indicator of a NE. Second, in our case, the temporal roles do not match the definition of the predefined NE classes, i.e. a temporal role is not a named entity, and as such the NER features do not necessarily capture the linguistic characteristics in text which are present in temporal roles, e.g. *head of state*.

Named Entity Disambiguation (NED) & Entity Linking (EL). In the case of NED, approaches [11, 15] are limited to resolving surface forms that point only to the predefined named entity classes (as shown above). As such, similar to the NER comparison, none of the NED systems is able to resolve cases like: "*head of state*" or "*U.S. President*". On the other hand, EL approaches [5] link surface forms to any Wikipedia page. One main disadvantage of existing EL systems is that they are trained on a specific Wikipedia snapshot, and as such the priors for a surface form linking to any of the Wikipedia pages are heavily biased towards the state of the Wikipedia snapshot. Furthermore, since EL link to any of the Wikipedia pages they do not resolve the temporal roles, but rather link to any matching Wikipedia page. That is, an example surface form "*U.S. President*" may link to the corresponding Wikipedia page, which is not disambiguated to the appropriate entity based on the context and temporal information present in the text. Finally, since in this paper we focus on the annotation of temporal roles in text, a comparison to named entity disambiguation techniques is not feasible.

Dictionaries and Gazetteers. One way to identify temporal roles in natural language is by making use of specific *cue phrases* or *surface forms* that are known to link to such roles. That is, given a dictionary of such cue phrases or surface forms for any given sentence to identify segments in the sentence (i.e. words) that link to a temporal role by a simple lookup in the dictionary. This constitutes an efficient approach and requires no computational efforts. However, an important prerequisite here is the dictionary generation process, which is domain dependent and varies given the genre of the considered text. GATE is a framework and graphical development environment that enables users to develop and deploy language engineering components [3]. One of the numerous components of the GATE pipeline is ANNIE (a Nearly-New Information Extraction System). This component contains a gazetteer for job titles, which to some extend might be considered as roles. Since this partly relates in terms of their objective, we consider ANNIE as a baseline for our temporal role annotation task.

Temporal Scope of Facts. In [17] the problem of determining the temporal scopes of facts in the form of RDF triples is investigated. In particular, the first step is to determine the time intervals in which a given fact is valid by exploiting evidence collected from the Web of Data and the Web. This approach addresses a more general setting in which the

Table 1: Distinction between *head role* and *role phrase* in a sentence.

Full Sentence	Head Role	Role Phrase
`The acting U.S President visited Pope Francis`	`President`	`The acting U.S. President`
	`Pope`	`Pope Francis`
`The World Soccer Champion lost the game yesterday`	`Champion`	`The World Soccer Champion`
`The Master of the Mint met Peter the Great`	`Master`	`The Master of the Mint`

temporal validity of a particular fact is evaluated, as e.g. "Barack Obama is the President of the United States". In this work, the assumption is that the facts are already extracted and structured in KBs, whereas in our case, we aim at extracting such structures from the text itself, i.e., identify the roles in a statement and later on disambiguate it to reference entities in a KB.

Standford's Title RegexNer [14] is able to extract job titles for people based on regular expressions. Our task does not restrict to the NE class `Person` only, thus, a direct comparison is not possible. Furthermore, this would require manual efforts to generate regular expressions, which would be subject to change for languages, text genre etc.

3. Problem definition

A *temporal role* can be defined in terms of different subject areas: linguistics, natural language processing, as well as ontology. The Cambridge Dictionary defines a role as *"the position or purpose that someone or something has in a situation, organization, society, or relationship"*[2]. Based on this definition, it can be induced that *a)* a role is a position, and *b)* anything (in particular named entities) can take over a role.

Moreover, since a *role* is not a named entity but refers to a named entity, it can be defined as a function which for a given *context* returns a unique *named entity*. In other words, a role might serve as anaphora or metaphor for a named entity. In this aspect, roles can be permanent, as e.g. *human*, or temporally restricted, as e.g. *student*, *CEO*.

In this paper, we focus especially on roles that specify *one* or only *a few* particular entities at a specific *point in time*, i.e. the inherent ambiguity of the role can only be resolved by (besides others) taking into account the *context* and *temporal information*. An example for such a temporal role is, "the *chancellor of Germany*". The particular individual referenced by that role depends on the given context present in the statement and the temporal information (i.e, the focus time point of a document or statement).

In this work, we distinguish between *head roles* and *role phrases*. A *head role* identifies the most general variant of a role with potentially high ambiguity. A *role phrase* further specifies the role with sufficient details for the potentially successful disambiguation. Table 1 contains several example sentences including the head roles and role phrases in those sentences. In *"the former U.S. President"* the term *President* is the head role, while the entire phrase, including potential further specifications, constitutes the role phrase. In this paper, with the term *role* we always refer to *head role*.

Let $S = (w_1, ..., w_n)$ be a natural language sentence that contains a finite number of words w_i with $1 < i < n$. Temporal role annotation refers to the identification of w_i in S, where w_i denotes a temporal role (head role). Furthermore, the minimal sequence $(w_j, ..., w_i, ..., w_k)$ has to be identified, with $j \leq i \leq k$ and $i, j, k < n$ such that $(w_j, ..., w_i, ..., w_k)$ specifies the role with sufficient details for a potentially successful disambiguation, i.e. any extension of the interval $(w_{j-1}, w_j, ..., w_i, ..., w_k)$ or $(w_j, ..., w_i, ..., w_k, w_{k+1})$ does not lead to further disambiguation.

[2] https://dictionary.cambridge.org/dictionary/english/role

Table 2: Feature list for the CRF-based role annotation approach

id	*Feature*	*Group*
f_1	w_i	Local Features
f_2	$POS(w_i)$	
f_3	$NER(w_i)$	
f_4	$startWithCapital(w_i)$	
f_5	$fullyInCapital(w_i)$	
f_6	$startOfSentence(w_i)$	
f_7	$Lemma(w_i)$	
$f_8 - f_{11}$	$w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}$	Contextual Features
$f_{12} - f_{15}$	$POS(w_{i-2}), POS(w_{i-1}), POS(w_{i+1}), POS(w_{i+2})$	
$f_{16} - f_{19}$	$NER(w_{i-2}), NER(w_{i-1}), NER(w_{i+1}), NER(w_{i+2})$	
f_{20}	$isInDic(w_i)$	External Features
f_{21}	$LocalGrammar(w_i)$	

4. Temporal Role Annotator

In this section, we present the features for our approach, which we employ in learning a linear-chain CRF. We use the model to annotate words or phrases with labels that are either "*role*" or "*no role*". For this task, we will extract features that exploit only the characteristics of the words themselves, the contextual features and further employ external features in the form of dictionaries whose entries correspond to surface forms that point to a temporal role.

In the following, we distinguish between two tasks of role annotation:

- Head Role Annotation (Sec. 4.1): Following the definition of *head roles* (Sec. 3) we propose a CRF-based approach to detect and annotate head roles in text.
- Role Phrase Annotation (Sec. 4.2): After the annotation of head roles, we use existing techniques to extract the parse tree of the sentence and to annotate the *role phrases*.

4.1. Head Role Annotation

For head roles annotation we propose an automated approach which addresses the problem as a *sequence classification* task. The choice of sequence classifiers is natural in our problem setting as the surface forms, which indicate a role, are influenced by the preceding and succeeding words in a sentence. Therefore, the relation between different segments in a sentence has a significant impact in determining the respective categorization, that is, a word (or a group of words) indicating a role or not. To capture such dependencies between the surface forms and the surrounding words in a sentence a linear-chain CRF model [12] has been chosen.

In the CRF-based role annotation approach, for a given sentence S from a document d, the sentence S is chunked into tokens consisting of words $S = (w_1, w_2, \ldots, w_n)$ and furthermore for each word w_i in S a set of *local* and *contextual* features is extracted, which capture intra-word dependencies and specific word characteristics. The complete list of features is shown in Table 2. Finally, the CRF-based sequence classifier for the given sequence of words from S predicts a label for each individual word indicating whether w_i is a role ($'R'$) or not ($'O'$).

In the following subsections the proposed features for the role annotation task are described in detail.

Features. This section describes the set of extracted features for a fragment w_i, which are used for training the linear-chain CRF for the role annotation task. We distinguish between two main groups of features depending on the way they are computed, that is *local* features, that make use only of a particular fragment in a sentence, and *contextual* features, which take into account the surrounding words of w_i.

Local Features. The local features take into account only the information from the individual fragments, that is, individual words from S. As local features are considered standard features that are used in natural language processing

tasks like part-of-speech tagging, named entity recognition etc. Often such features provide important information depending on the task at hand. For temporal role annotation, a word indicating a role belongs to either the POS tags NN (noun) or NNP (proper noun) (see Example 1 below). Other local features that help to detect roles are summarized and briefly explained below.

- f_1**:** The current token, w_i.
- f_2**:** The POS tag of the current token computed using Stanford CoreNLP [14].
- f_3**:** The named entity class of the current token [14]. If the current token is not labeled with any NER class then the value of the feature is *null*.
- f_4**:** A boolean flag denoting if the current token starts with a capital letter.
- f_5**:** A boolean flag denoting if the current token is in all caps.
- f_6**:** A boolean flag denoting if the current token is the first token of the sentence.
- f_7**:** Lemmatization of the current token.

Example 1. Given the following sentence *"The president of France gave a speech yesterday."* and $i = 2$ the feature value list is: { f_1="president", f_2="noun", f_3="null", f_4=false, f_5=false, f_6=false, f_7="president" }.

However, from the examples shown in Table 1 and the example above, it is obvious that temporal roles cannot be detected solely based on local features. This is because words indicating a role often belong to the POS tag NNP, which might be part of a named entity. However, the performance of the NER system is not always perfect and thxcerefore many named entities can be missed. As such, feature f_3 alone will not be able to distinguish whether a detected named entity is also a role. Therefore, additional contextual features are considered that exploit the dependencies among surrounding words in S.

Contextual Features. In this feature group the dependencies of a word w_i with its surrounding words w_{i-1} and w_{i+1} are captured. For example, given the word w_i=*"president"* and the succeeding word w_{i+1}=*"of"*, this represents a strong signal for the sequence classifier indicating that w_i is a role. In the following the remainder of the features of the *contextual* feature group are explained in detail.

- $f_8 - f_{11}$**:** Represent the surrounding tokens to the token w_i in a window of two preceding and succeeding words. The surrounding context has proven to be an important feature in similar NLP sequence classification tasks. For an ambiguous word w_i its context provides discriminatory information to determine whether the word indicates a role or not. Consider the following sentence, *"The king in chess is the most important piece."* with w_i=*"king"*. In this particular case, the word w_{i+2}=*"chess"* plays a crucial role in the decision process.
- $f_{12} - f_{19}$**:** Contrariwise to features f_8–f_{11}, which use the actual token values as features, the syntactical features consisting of the POS tags of the surrounding tokens of w_i are considered. These features exploit the commonalities in the language structure for role annotation, as e.g., NN IN NNP (*"president of France"*) is a frequent pattern that indicates a role. Similar to POS tags for the context of w_i, the named entity classes are considered for the surrounding words.

In summary, the combination of local and contextual features provide high generalization power over previously unseen roles in different textual corpora. Through the local features the syntactic and semantic information of the tokens is captured, and this in combination with the contextual features captures patterns in language being a combination of both syntactic and semantic information, e.g. *the president of* <*Location*> from the example above, which can be used to correctly detect the role *"president"*.

External Features. As external features are considered the features that do not extract information from the current word w_i neither from the surrounding context, like the previous feature groups, but they use knowledge coming from external sources to further assist in the correct detection of a role.

- f_{20}**:** A boolean flag indicating if a word in the token sequence of S is part of a given dictionary.
  To the best of our knowledge there is no dictionary dedicated to list all temporal roles. Thus, in this paper we automatically construct a dictionary dedicated to the special case of temporal role annotation.

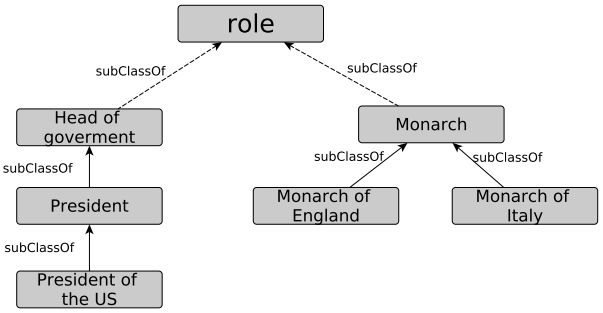

Fig. 2: Subtree of the Wikidata class hierarchy for the class *"role (Q214339)"*.

Dictionary Creation. In this process, an important point of consideration is the coverage as well as the temporal aspect of the surface forms in the dictionary, that fit our problem setting. For this purpose, the Wikidata knowledge graph was considered, a widely used source of information with constantly evolving data about real-world entities and concepts.
According to its definition, the Wikidata class *role (Q214339)* is the root class in the Wikidata subtree of *role-classes* that contain all possible roles one can have in terms of *"rights, obligations, beliefs, and norms expected from an individual that has a certain social status"*. Fig. 2 shows a fraction of the aforementioned subtree. In this case, the class labels serve as surface forms that map to a role, and therefore are considered as *candidates* for addition into the dictionary. Next, to ensure the temporal aspect of the roles, consequentially, the surface forms indicating a role, the instances of the class *role (Q214339)* as well as its subclasses are determined. To identify those instances, time-dependent properties are considered, such as, e.g., *"replaces"* or *"replaced by"*, as shown in Listing 1. Finally, the candidate class labels are added to the dictionary labels fulfilling the criteria of the SPARQL query depicted in Listing 1, which identifies classes with temporal roles.
We further improved the completeness of the dictionary by taking into account the anchor text from the equivalent English language Wikipedia articles corresponding to the role class *"role (Q214339)"* and its subclasses[3]. This step contributes a significant amount of surface forms in the generated dictionary. However, the anchor texts in Wikipedia are often arbitrary and the granularity of the anchor texts vary from case to case, since such links are added manually by Wikipedia editors. To avoid the introduced errors and redundancies, the dictionary is subject to subsequent filter and normalization steps, which include duplicates removal, removal of special characters and punctuation marks, as well as the removal of role phrases from the dictionary.
In the final state, our dictionary consists of more than 200 unique surface forms that point to our target head roles.

```
SELECT DISTINCT ?role ?roleLabel WHERE {
  ?role instanceOf*|subClassOf* wd:Q214339 .
  ?role label ?roleLabel.
  ?person positionHeld ?roleStatement.
  ?roleStatement positionHeld ?role.
  ?roleStatement replaces|replacedby ?differentRoleHolder.
}
```

Listing 1: SPARQL query Q_1

- f_{21}**:** A boolean flag indicating if a word matches to a Local Grammar (LG). Local grammars are handmade rules used to identify temporal roles in text and are one means of representing the contextual rules of the linguistic approach [9]. For example, *ROLE IN(POS) LOCATION(NER)* is able to detect phrases such as "President of Iran" or "King of Egypt". In this paper we proposed five handmade local grammars which are listed in Table 3:

[3] In Wikidata, the role classes and its subclasses provide links to the corresponding Wikipedia articles.

Table 3: Local Grammars

Local Grammar	Samples
`ROLE IN(POS) LOCATION(NER)`	President of Iran, King of Egypt
`ROLE PERSON(NER)`	President Obama, Pope Francis
`DT(POS) ROLE`	The King, The Pope
`ROLE IN(POS) ORGANIZATION(NER)`	CEO of Apple
`ORGANIZATION(NER) ROLE`	Google CEO

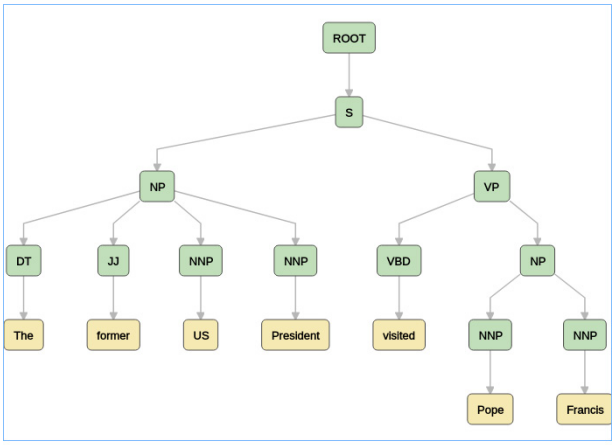

Fig. 3: Constituency parse tree for the sentence "The former US President visited Pope Francis".

4.2. Role Phrase Annotation

Up to now, the statistical Learning approach (Sec. 4.1) is able to correctly annotate *head roles* in a natural language text. However, through this approach we cannot annotate *role phrases*, as they were defined in Sec. 3. To be able to solve this task correctly, further semantics has to be taken into account, as e.g., constituency parse trees. A parse tree is an ordered, rooted tree that represents the syntactic structure of a textual fragment according to some context-free grammar. It is able to detect a noun phrase (abbreviated as NP), which is a phrase that has a noun as its head. In the scope of this work, as *role phrase* the longest (nearest to the root) NP is selected that contains a role. Fig. 3 depicts the parse tree for the example sentence *"The former US President visited Pope Francis."*. In this example *"The former US President"* and *"Pope Francis"* are both considered as *role phrases*.

5. Experimental Setup

This section outlines the experimental setup for the evaluation of the temporal role annotation approach with the competitors for this task. The data and the source code of the proposed approaches are publicly available here[4]. All the experiments conducted in this paper have been executed on a machine with Core i7 CPU and 16 GB RAM.

Table 4: Automatically Generated Dataset Statistics

Sample	Number of Sentence	Number of Word Tokens
Positive	199141	6 M
Difficult Negative	48135	1.5 M
Easy Negative	151006	3.3 M

[4] https://github.com/ISE-AIFB/RoleTagger

Table 5: Manually curated ground-truth statistics.

Role group	President	Pope	Monarch	CEO
#True Positive Tags	47	74	33	19
#False Positive Tags	6	37	30	3

5.1. Ground-Truth Dataset

To test the proposed approach for temporal roles annotation we created two ground-truth datasets. Both datasets examine four major categories of temporal roles. Those categories are *President*, *Monarch*, *Pope* and *CEO*. We focused on the four mentioned roles as they represent highly ambiguous cases for the later steps of disambiguation. Other roles can be tested, and we plan to do so in the future steps, but we believe that those are representative cases to evaluate our approach.

Distant Supervision Ground-Truth. This section describes the process of generating the ground-truth dataset based on distant supervision. To guide the distant supervision process, the following input information is necessary:

- E – List of Wikipedia articles (of type `Person`) that are instances of the classes as result of the SPARQL query in Listing 1.
- D – Dictionary of surface forms that indicate temporal roles (cf. Sec. 4.1)
- T – Induced category subtree of all classes returned by the query in Listing 1.

The aim is to extract the following three sets of samples: (i) *positive samples* (sentences that contain a temporal role), (ii) *easy negative samples*, which indicate samples that do not contain any surface form that links to a temporal role, and (iii) *difficult negative samples* with textual fragments present in the constructed dictionary but not referring to any temporal role.

The data gathering process starts with a Wikipedia article P and sentence $S \in P$. S is considered to be a *positive* sample if its *anchor text* contains textual fragments present in D and additionally it anchors to a Wikipedia article P' that is part of E. In case P' does not belong to E, however, if there exist at least one category of P' that is part of the category subtree T, S is considered a positive sample. Finally, in the last case, S is considered a positive sample, if it contains textual fragments (without any anchor) present in D and P that is part of the entity set E.

In the case where S does not contain any textual fragment in dictionary D, then it is considered as an *easy negative*.

Otherwise, if S contains an anchor text where at least one fragment is present in D, whereas it links to a Wikipedia article P' that is not part of E nor any of its categories are part of the category subtree T, then it is considered as a *difficult negative* sample.

Finally, to ensure the quality of our constructed ground-truth, we manually checked a random sample of more than 700 instances and assessed if our labelling through distant supervision as *positive* and *difficult negative* is correct. The quality in these two cases is 97%, which presents a high quality and large scale dataset for this task.

Manually Curated Ground-Truth. In this case, we manually creates a ground-truth dataset by sampling sentences from The New York Times news corpus [18]. The main aim of extracting this dataset is to assess how well our approach generalizes on different genres of text. Similarly, here too, the samples contain instances of the following role groups: *President*, *Pope*, *Monarch* and *CEO*.

The dataset consists of a total of 200 sentences which we sample by extracting those sentences that contain one of the surface forms from our dictionary. From the resulting sample, 64% are instances that contain a role, whereas the remainder of 36% consists of instances that are labelled as difficult negative cases, where there are surface forms that are present in our dictionary but do not represent roles. The dataset is available for download here[5]. General statistics for the generated ground-truth dataset including sample sizes for the different role groups are shown in Table 5.

[5] https://github.com/ISE-AIFB/RoleTagger/tree/master/groundtruth

5.2. Learning Framework

In this study CRFSuite [16] has been applied, which is an open source implementation of CRFs. For parameter optimization, Averaged Perceptron is utilized with 200 iterations. For each token in a sentence it outputs either ($'R'$) or ($'O'$) depending on weather it holds a temporal role in the annotated dataset or not.

5.3. Train/Test

The CRF-based model is evaluated based on the following strategies:

- First, for the distant-supervised ground-truth a 5-fold cross validation (CV) is performed to assess the robustness of the approach. The exhaustive nature of this dataset consisting of 6 million positive tokens and 4.8 million negative tokens allows to provide conclusive results for the role annotation task.
- Second, to account for possible false positives and negatives in the distant-supervision ground-truth dataset, the CRF-based model is trained on the distant-supervision dataset, and tested on the manually curated ground-truth dataset, consisting of a total of 200 sentences. This presents a highly rigorous evaluation as we train the models on samples extracted from Wikipedia, and test it on samples from a different genre and domain such as news articles from the New York Times corpus.

5.4. Baselines

The proposed CRF-based temporal role annotation approach is tested against the following baselines:

- **Our Dictionary:** a dictionary generated by considering Wikipedia and Wikidata anchor-texts with 210 unique roles (Sec. 4.1 *Dictionary creation*).
- **ANNIE job title gazetteer [3]:** a dictionary containing 1.567 job titles.

6. Results and Discussion

This section reports and discusses the evaluation results for the role annotation task.

Performance. Table 6 shows the evaluation results for our approach based on 5-fold cross validation for the ground-truth that we generate through distant supervision. We compare our approach against a simple baseline in this case, where we annotate surface forms in statements as having a temporal role if they match any of the entries in our dictionary. While it is intuitive for the baseline in Table 6 to have perfect recall, it is interesting to note on the other hand that our approach achieves a nearly perfect recall with 97%. In terms of precision, we have 12% relative improvement over the baseline, which presents a significant improvement[6]. Similar in terms of F1 score, we outperform the baseline approach with nearly 6% relative improvement.

In detail, with the dictionary we are able to filter out the *easy negative samples* but not the samples with ambiguous surface forms, which may refer to a person name or a role (as e.g., "Pope"). This shows the advantage of our approach, which can accurately distinguish between the surface forms that point to a role and the difficult negative samples that contain a surface form from the dictionary but do not contain any role. Table 8 presents the feature ablation, explaining in detail the performance of our approach based on the different feature groups.

In the following, we show how robust and how well our approach does generalize if we train in the distant supervision ground-truth and test on the manually curated ground-truth, which consists of sentences of a different genre.

Robustness. Table 7 shows the evaluation results for all role annotation approaches under comparison, where we train our model in the distant supervision dataset and test on the manually curated ground-truth. In terms of precision, the CRF-based *Role Annotator* presented in this paper achieves the best results with a relative improvement of over 14%

[6] We tested using t-test statistic with $p < 0.05$.

Table 6: Evaluation results on the distant supervision ground-truth.

Methods	Precision	Recall	F_1
Our Dictionary	0.82	1.00	0.90
CRF-based Role Annotator (5-fold CV)	**0.94**	0.97	**0.95**

compared to the dictionary-based role detection approach. In terms of recall, the dictionary based approach achieves perfect recall, which is a 7% relative increase compared to the CRF-based approach. This is intuitive, as the dictionary is used in selecting the samples from the NYT corpus by considering only sentences which match one of the surface forms from our dictionary.

A post-hoc analysis of the evaluation results reveals that all errors of the dictionary-based model represent *difficult negative samples*. For example, the dictionary-based cannot differentiate the word "Pope" in "Alexander Pope" and "Pope Francis", i.e. whether it is a family name or actually a temporal role, respectively. While the CRF-based approach through the encoded features learns patterns like "<PERSON> Pope", which help to distinguish between words representing a role or not.

The baseline ANNIE, which represents a simple gazetteer that uses a predefined list of job-titles, achieves an F_1 score of 62%, which in comparison to the results of our CRF-based approach, we achieve a 24% relative improvement in terms of F_1 score.

Table 7: Evaluation results on the manually-curated ground-truth.

Methods	Precision	Recall	F_1
Our Dictionary	0.66	**1.0**	0.79
ANNIE Gazetteer	0.55	0.72	0.62
CRF-based Role Annotator	**0.80**	0.93	**0.86**

Feature Ablation. Table 8 shows the ablation results for the different feature groups for the CRF-based Role Annotator. It is interesting to note that the local features provide the highest performance gain, given that they use only information about the current tokens in a sequence. As such, they represent a feature which generalizes well, since it does not require external knowledge as is the case for the external feature group. Whereas, in the case of contextual features as they consider longer sequences, the features are more sparse and thus the recall score is significantly lower in comparison to other feature groups. In the case of external features, apart from its manual efforts and its limited applicability outside the domain of the sources used to generate the dictionary itself, the performance of the CRF model is comparable to the local feature group in terms of the achieved F_1 score.

The insights here are that considering the combination of these features yields optimal performance as it learns to detect surface forms that consist of more than one token, which is impossible for local or external features alone without the contextual feature group. Finally, to be more fair to the model, we performed a last feature ablation experiment. This time the f_{20} feature uses only the 75% of the initial dictionary, to measure the impact of the given dictionary.

Table 8: Feature Ablation (5-fold Cross Validation)

Feature Group	Precision	Recall	F_1
Local Features	0.90	0.95	0.92
Contextual Features	0.72	0.27	0.39
External Feature	0.82	1.00	0.90
All Features	0.94	0.97	**0.95**
All Features (25% Dictionary size reduction)	0.94	0.92	0.93

7. Conclusion and Future Work

In this paper we have presented a structured prediction approach that relies on a linear-chain CRF for annotating temporal roles in natural language text. For the model we exploit local, contextual, and external features extracted from existing knowledge bases like Wikidata. Furthermore, we have automatically compiled a high quality dictionary of surface forms that link to temporal roles. The approach has been evaluated on two ground-truth datasets. First, we have automatically constructed a high quality dataset through distant supervision by relying on our dictionary. Our model achieves an F_1 score of 95% in the role annotation process, by outperforming significantly our baseline approaches. Next, to assess how well our model generalizes, we have trained on the distant supervision dataset and evaluated on a manually curated ground-truth consisting of a small sample of sentences from the NYT corpus, where is shown that our model can annotate with high accuracy and outperform the baselines.

As future work, we plan to continue with the subsequent step, which resolves the annotated surface forms with temporal roles to the respective named entities from a reference knowledge base by taking into account the context including temporal information.

References

[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. Dbpedia: A nucleus for a web of open data, in: The semantic web. Springer, pp. 722–735.

[2] Baker, C.F., Fillmore, C.J., Lowe, J.B., 1998. The berkeley framenet project, in: Proceedings of the 17th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics. pp. 86–90.

[3] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, in: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).

[4] Dutta, A., Meilicke, C., Stuckenschmidt, H., 2015. Enriching structured knowledge with open information, in: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland. pp. 267–277. URL: https://doi.org/10.1145/2736277.2741139, doi:10.1145/2736277.2741139.

[5] Ferragina, P., Scaiella, U., 2012. Fast and accurate annotation of short texts with wikipedia pages. IEEE Softw. 29, 70–75.

[6] Finkel, J.R., Grenager, T., Manning, C., 2005. Incorporating non-local information into information extraction systems by gibbs sampling, in: ACL.

[7] Galárraga, L., Heitz, G., Murphy, K., Suchanek, F.M., 2014. Canonicalizing open knowledge bases, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, New York, NY, USA. pp. 1679–1688. URL: http://doi.acm.org/10.1145/2661829.2662073, doi:10.1145/2661829.2662073.

[8] Gildea, D., Jurafsky, D., 2002. Automatic labeling of semantic roles. Computational Linguistics 28, 245–288.

[9] Gross, M., 1999. A Bootstrap Method for Constructing Local Grammars, in: Bokan, N. (Ed.), Proceedings of the Symposium on Contemporary Mathematics. University of Belgrad, pp. 229–250. URL: https://halshs.archives-ouvertes.fr/halshs-00278319.

[10] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G., 2011a. Robust disambiguation of named entities in text, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 782–792. URL: http://dl.acm.org/citation.cfm?id=2145432.2145521.

[11] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G., 2011b. Robust disambiguation of named entities in text, in: EMNLP, pp. 782–792.

[12] Lafferty, J., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Morgan Kaufmann. pp. 282–289.

[13] Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data .

[14] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D., 2014. The Stanford CoreNLP natural language processing toolkit, in: Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60.

[15] Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C., 2011. Dbpedia spotlight: Shedding light on the web of documents, in: Proceedings of the 7th International Conference on Semantic Systems, ACM, New York, NY, USA. pp. 1–8.

[16] Okazaki, N., 2007. CRFsuite: a fast implementation of conditional random fields (CRFs).

[17] Rula, A., Palmonari, M., Ngomo, A.N., Gerber, D., Lehmann, J., Bühmann, L., 2014. Hybrid acquisition of temporal scopes for RDF data, in: The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings, pp. 488–503.

[18] Sandhaus, E., 2008. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia .

[19] Suchanek, F.M., Kasneci, G., Weikum, G., 2007. Yago: a core of semantic knowledge, in: Proceedings of the 16th international conference on World Wide Web, ACM. pp. 697–706.

[20] Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S., 2007. Textrunner: open information extraction on the web, in: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Association for Computational Linguistics. pp. 25–26.