Collective Information Extraction using First-Order Probabilistic Models

Slavko Žitnik[*]
University of Ljubljana
Faculty of Computer and
Information Science
Ljubljana, Slovenia
slavko.zitnik@fri.uni-lj.si

Lovro Šubelj
University of Ljubljana
Faculty of Computer and
Information Science
Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

Dejan Lavbič
University of Ljubljana
Faculty of Computer and
Information Science
Ljubljana, Slovenia
dejan.lavbic@fri.uni-lj.si

Aljaž Zrnec
University of Ljubljana
Faculty of Computer and
Information Science
Ljubljana, Slovenia
aljaz.zrnec@fri.uni-lj.si

Marko Bajec
University of Ljubljana
Faculty of Computer and
Information Science
Ljubljana, Slovenia
marko.bajec@fri.uni-lj.si

ABSTRACT

Traditional information extraction (IE) tasks roughly consist of named-entity recognition, relation extraction and coreference resolution. Much work in this area focuses primarily on separate subtasks where best performance can be achieved only on specialized domains.

In this paper we present a collective IE approach combining all three tasks by employing linear-chain conditional random fields. The usage of probabilistic models enables us to easily communicate between tasks on the fly and error correction during the iterative process execution. We introduce a novel iterative-based IE system architecture with additional semantic and collective feature functions.

Proposed system is evaluated against real-world data set, introduced in the paper, and results are better over traditional approaches on two tested tasks by error reduction and performance improvements.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; I.5.4 [**Pattern recognition**]: Applications—*Text processing*

General Terms

Algorithms, Experimentation, Languages

Keywords

information extraction, Conditional Random Fields text tagging

1. INTRODUCTION

Machine understanding of textual documents, needed at IE, has been challenging since early computer-era. IE lies between information retrieval systems [12], which finds documents related to user's requirements, and text understanding systems that attempt to analyze text and extract their semantic contents. Early IE methods were naive and rule based, then (semi-) automatic approaches of wrapper generation, seed expansion or rule induction were developed and recently machine learning techniques gained popularity. In contrast to standard multi-label and regression classifiers, sequence taggers such as Hidden Markov Models, Maximum Entropy Models and Conditional Random Fields (CRF) have become the most successful. Especially the latter, which supports rich definition of feature functions.

Main IE tasks consist of named entity recognition (NER) - (e.g. extraction of names, locations, organizations), relation extraction (RE) - (e.g. identification of connection types among entities) and coreference resolution (COREF) - (e.g. clustering of mentions to an entity). A vast majority of research focuses only on one IE subtask or a pipeline of them not interconnecting them together.

In this paper we propose a collective IE algorithm that iteratively combines all three subtasks. We employ a linear-chain CRF algorithm for every subtask and present additional iterative and semantic features. The use of the same learning techniques enables us to easily use a subtask output labelings as feature inputs for others. In addition to labeling tasks we introduce an entity resolution technique for coreferent mentions matching and merging. Furthermore, during clustering new semantic attributes with coreferent entity values are appended to existing ones, used by CRF's feature functions.

The rest of the paper is organized as follows. Section 2 gives

[*]Author is also actively engaged in industry research work at Optilab d.o.o., Teslova 30, SI-1000 Ljubljana

a brief review of related work, focusing mainly on collective IE. Next, a dataset labeling for all three subtasks is presented, followed by CRFs definition, presentation of novel collective algorithm and introduction of new feature functions. Some preliminary results on real-world dataset are discussed in Section 4, and lastly Section 5 concludes the paper and reveals further work.

2. RELATED WORK

As mentioned earlier, a majority of research focuses on each subtask separately. The latest research results show the entity extraction problem is quite well solved as best methods achieve 90% F-score on general datasets [11, 2]. On the opposite side, at RE and COREF state-of-the-art methods achieve roughly about 70% F-score [6, 5].

The use of semantic data has also been introduced for IE problems. The general semantic source is Wordnet which contains groups of words with similar meaning - synsets. More exact way of using semantics is an ontology which is defined as explicit specification of conceptualization and provides schema, rules and instances. Recently, ontology-based IE emerged as a subfield of IE [7] because schema, rules and data interoperability can be sufficiently used and merged.

The term collective information extraction was to our knowledge for the first time used on iterative NER exploiting mutual influence between possible extractions [1]. Later Ontology-based IE [8] that in a cyclic process combines NER and RE with knowledge integration using an ontology was proposed. The system was completely rule based, but it pointed the right direction. The most recent system, Felix [10], is a general IE system based on logical and statistical rules that use Markov Logic Networks. The authors focused on scaling it to large datasets and definition of generally applicable rules. They tested their iterative method on NER and COREF tasks, but the system needs input of evidence, examples and rules as input.

Early work in IE was driven by challenges at MUC[1], CoNLL[2] conferences and ACE[3] program. Along with tasks, the datasets were provided and they are still used for system evaluations.

Conditional Random Fields [4], a sequence modeling framework, have been used on various sequence labeling tasks. At proper text labeling and feature induction they were successfully applied to the task of NER [2], RE [6] and COREF [13]. The latter is often accompanied by clustering methods as coreferent mentions need to be merged.

3. COLLECTIVE IE METHOD

In this section we introduce dataset representation, present used methods and propose a system for collective IE.

3.1 Representation

We treat the tasks of NER, RE and COREF as sequence labeling tasks.

Let $\overline{x}^{k_i} = \{x_1^{k_i}, x_2^{k_i}, ..., x_n^{k_i}\}$ denote the sequence of observable tokens. Index k_i stands for input words w_i or additional attributes such as part-of-speech (POS) tags, phrase boundaries, entity cluster inclusion or pre-calculated values. Each observable sequence is associated with corresponding labeling sequence $\overline{y}^{l_i}$ where $l_i \in \{NE, REL, COREF\}$ is defined for named entity, relation and coreference label tags.

We use common IOB notation for all three types of sequence labeling. Tags starting with "B-" denote start of a label type, "I-" the successor of the same type and "O" other types. An example label tag set for person named entities is {B-PER, I-PER, O}. For relations we use labels {B-REL, I-REL, O}. Coreference mentions are labelled using set {B-COREF, I-COREF, O}. Here, tag is labeled as "I-COREF" if and only if it is coreferent with previous (possible distant) tag, labelled as "B-COREF".

Our problem is now finding the most probable labelings $\hat{y}^l$ for each of defined subtasks.

3.2 Conditional Random Fields

A Conditional Random Fields (CRFs) [4] are discriminative models and model a single joint distribution $p(\overline{y}|\overline{x})$ over the predicted sequence $\overline{y}$ conditioned on $\overline{x}$. Observable sequence $\overline{x}$ typically contains also a number of attributes that can be used when modeling feature functions. Used training labels $\overline{y}$ relative to position i inside feature functions f_j define the structure of model which can in general be arbitrary.

At CRFs training we are looking for a weight vector w that assigns best possible labeling $\hat{y}$ given $\overline{x}$ for all training examples:

$$\hat{y} = \arg\max_{\overline{y}} p(\overline{y}|\overline{x}; w), \tag{1}$$

using conditional distribution

$$p(\overline{y}|\overline{x}; w) = \frac{\exp(\sum_{j=1}^{J} w_j \sum_{i=1}^{n} f_j(\overline{y}, \overline{x}, i))}{Z(\overline{x}, w)} \tag{2}$$

Vector w contains a real number for every possible input (J inputs) to f_j.($Z(\overline{x}, w)$ is a normalization constant over all possible labelings of $\overline{y}$). When distance between two addressing labels inside feature functions is long, exact inference is intractable due to exponential number of partial sequences and thus approximate algorithms must be used. We therefore use feature functions that depend only on single label (y_i) and two consecutive labels (y_{i-1}, y_i). This type of CRF is also known as linear chain CRFs (LCCRFs) which underlying graphical structure forms a chain and have been rather successful in IE tasks. Using LCCRFs, training and inference can be easily solved using forward–backward method and Viterbi algorithm.

3.3 Collective approach

We propose a collective IE algorithm combining tasks of NER, RE and COREF. A high level implementation of iterative training and labeling algorithm is shown as Algorithm (1) and Algorithm (2). The approach should give better results because it takes into account intermediate labelings of other subtasks in an innovative manner.

The input parameters for training Algorithm (1) are sen-

[1] Message Understanding Conference

[2] Conference on Computational Natural Language Learning

[3] Automatic Content Extraction

tences, tokenized by words with additional attributes $\overline{x}^k$, true named entities, relations and coreferences labelings denoted as $\overline{y}^l$ and number of maximum possible iterations. The final training result is a 3-tuple of trained classifiers for each task. These classifiers can be used independently, but are trained to be used by Algorithm (2) to get best results. During each learning iteration in Algorithm (1), feature function vectors are initialized and then classifiers are independently learned. After that we update/create additional attributes for next iteration of training. At that step we perform collective entity resolution using attribute, relational and semantic similarity measures as proposed and evaluated in [14]. As a result we get clusters of coreferent entities which values are used at initialization of feature vectors in next iteration. Similarly we provide additional attributes by tagging the input sequence using latest classifiers. For example, let have an input sequence: "*John has left ACME. ... When he worked at ACME as a student.*" In iteration i ACME was not recognized as a company in first sentence, but all others were correctly labeled and pronoun *he* was merged with *John*. In the next iteration feature functions can use distant relation *workedAt* **Company** which results in correct first sentence labeling. Iterating ends when classifier's labelings over iterations converge or maximum number of iterations is achieved. We will empirically define *maxIter* and *prevScoreDiff* in further work with all three subtasks.

Algorithm 1 Collective IE Training

```
Input: $\overline{x}^k$, $\overline{y}^l$, maxIter
Output: classifiers (cNE, cREL, cCOREF)
Initialize coref. clusters as $C = \emptyset$
$i \leftarrow 0$
while $i <$ maxIter and prevScoreDiff() $< \varepsilon$ do
    Initialize feature functions
    cNE $\leftarrow$ LCCRF($\overline{x}^k, \overline{y}^{NE}$)
    cREL $\leftarrow$ LCCRF($\overline{x}^k, \overline{y}^{REL}$)
    cCOREF $\leftarrow$ LCCRF($\overline{x}^k, \overline{y}^{COREF}$)
    $C \leftarrow$ entityResolution($\overline{x}^k, \overline{y}^l$)
    $\overline{x}^{I\_NE} \leftarrow$ cNE.tag($\overline{x}^k$)
    $\overline{x}^{I\_REL} \leftarrow$ cREL.tag($\overline{x}^k$)
    $\overline{x}^{I\_COREF} \leftarrow$ cCOREF.tag($\overline{x}^k$)
    $i \leftarrow i + 1$
end while
return (cNE, cREL, cCOREF)
```

Algorithm (2) introduces iterative labeling and is very similar to training algorithm. Only feature initialization, tagging and coreference clustering is used until there are no labeling differences over two sequential iterations or maximum number of them is reached.

3.4 Features

The selection of feature functions is an essential step for successfully training CRF classifiers.

We use proposed NER feature functions by Manning et. al. [2], RE-specific features proposed by Li et. al. [6] and COREF-specific features by McCallum et. al. [13] and Ng. and Cardie [9]. The union of all features across tasks represents word, text preprocessing (i.e. POS tags, lemmas, Parse trees) and word shape features.

Algorithm 2 Collective IE Labeling

```
Input: $\overline{x}^k$, (cNE, cREL, cCOREF), maxIter
Output: labelings and coreference clusters
Initialize coref. clusters as $C = \emptyset$
$i \leftarrow 0$
while labelingsChanged() and $i <$ maxIter do
    Initialize feature functions
    $\overline{x}^{I\_NE} \leftarrow$ cNE.tag($\overline{x}^k$)
    $\overline{x}^{I\_REL} \leftarrow$ cREL.tag($\overline{x}^k$)
    $\overline{x}^{I\_COREF} \leftarrow$ cCOREF.tag($\overline{x}^k$)
    $C \leftarrow$ entityResolution($\overline{x}^k, \overline{y}^l$)
    $i \leftarrow i + 1$
end while
return ($\overline{x}^{NE}$, $\overline{x}^{REL}$, $C$)
```

Table 2: A distribution of BIO tags following proposed representation in Section 3.1

Type	B-	I-	O
Named entities	293	233	5508
Relations	32	24	5978
Coreferences	274	249	5511

In Table 3.4 we introduce additional iterative and semantic feature functions. It is worth mentioning that some local attributes are equivalent to long-distance that can be modeled as arbitrary structured CRF and are here a result of entity resolution.

4. EVALUATION AND RESULTS

We conducted some preliminary analysis of part of proposed method. We employed only tasks of NER and RE to show the results are promising and it is worth to build the whole system.

We tested methods on real-world news dataset in Slovene language which is publicly available [4]. The topic of the articles is mainly political. It contains 6034 word tokens within 285 sentences. The tokens are annotated according to proposed representation in Section 3.1 (their distribution is shown in Table 2). We additionally lemmatized and POS-tagged the whole corpus using slovene POS tagger [3]. Slovene is morphologically complex language and therefore around thousand different POS tags exist. Within the dataset there are 315 distinct POS tags labeled.

For evaluation purposes we implemented linear-chain CRFs using stochastic gradient ascent learning algorithm and Viterbi for decoding with rich feature function API that supports arbitrary implementation. Whole implementation with additional broader framework features is available online [5].

In Table 3 we show achieved MAF (macro averaged F-score) measure when training independently, using traditional "pipeline" approach and by employing part of proposed collective algorithm. In macro-averaging, F-measure is computed locally over each category first and then the average over all categories is taken. All approaches use feature functions

[4] `http://zitnik.si/mediawiki/index.php?title=File:Rtvslo_dec2011.tsv`

[5] `https://bitbucket.org/szitnik/iobie/`

Table 1: Linear-chain feature function templates. i indicates current position and j offset relative to i. Functions depending only at one label generate $\#labels$ features and $\#labels^2$ features for depending on two consecutive labels.

Feature Function Description	Feature Femplate	Example
single cluster relation	$c_i(-1 \leq i \leq 1)$	c_i is *works at*
single cluster entity tag	$c_i(-1 \leq i \leq 1)$	c_i is *I-PER*
single cluster entity name	$c_i(-1 \leq i \leq 1)$	c_i is *Chuck Norris*
single previous iter NE tag	$n_i(-2 \leq i \leq 2)$	n_i is *B-PER*
single previous iter REL tag	$r_i(-2 \leq i \leq 2)$	r_i is *I-REL*
single previous iter COREF tag	$co_i(-2 \leq i \leq 2)$	co_i is *B-COREF*
two consecutive entity tags at coreferences	$c_{i+j-1}\&c_{i+j}(-1 \leq j \leq 2)$	c_{i+j-1} is *I-ORG* and c_{i+j} is *B-PER*
two consecutive iter NE tags	$n_{i+j-1}\&n_{i+j}(-1 \leq j \leq 2)$	n_{i+j-1} is *O* and n_{i+j} is *B-ORG*
two consecutive iter REL tags	$r_{i+j-1}\&r_{i+j}(-1 \leq j \leq 2)$	r_{i+j-1} is *B-REL* and r_{i+j} is *I-REL*
two consecutive iter COREF tags	$co_{i+j-1}\&co_{i+j}(-1 \leq j \leq 2)$	co_{i+j-1} is *O* and co_{i+j} is *B-COREF*

Table 3: Comparison of macro averaged F-score on real world dataset by independent, pipeline and collective learning

	Entity Recognition	Relation Extraction
Independent	.57	.71
Pipeline	.57	.72
Collective	**.59**	**.74**

leveraging learning label, upper case of one and two consecutive words, prefixes and suffixes of length two and three and POS, words and lemma features. At collective approach, additional iterative feature functions for NER and RE are used. Results show the collective approach outperforms other two in both labeling tasks.

5. CONCLUSIONS AND FUTURE WORK

Paper proposes a collective information extraction algorithm, which combines tasks of named entity recognition, relation extraction and coreference resolution. We introduce iterative training and labeling algorithm, present new iterative feature functions and show preliminary experimental results which show improvements over traditional approaches.

Future work will include implementation of the whole proposed iterative system, modeling of new features and introduction of parallelization algorithms. After that we will be able to compare results to others on larger datasets. Ontologies will be incorporated into the algorithm to use additional manual patterns, constraints and to better connect with entity resolution module.

6. ACKNOWLEDGMENTS

The work has been supported by the Slovene Research Agency ARRS within the research program P2-0359 and part financed by the European Union, European Social Fund.

7. REFERENCES

[1] R. Bunescu and R. J. Mooney. Collective information extraction with relational markov networks. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, 2004.

[2] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, 2005.

[3] M. Grčar, J. Rupnik, M. Juršič, and S. Krek. Slovene pos tagger. http://označevalnik.slovenščina.eu/ Vsebine/ Sl/ProgramskaOprema/ Meta.aspx.

[4] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.

[5] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the CoNLL-2011 Shared Task*, 2011.

[6] Y. Li, J. Jiang, H. Chieu, and K. Chai. Extracting relation descriptors with conditional random fields. pages 392–400, Thailand, 2011. Asian Federation of Natural Language Processing.

[7] L. McDowell and M. Cafarella. Ontology-driven information extraction with ontosyphon. *The Semantic Web-ISWC 2006*, page 428–444, 2006.

[8] C. Nedellec and A. Nazarenko. Ontologies and information extraction. *CoRR*, abs/cs/0609137, 2006.

[9] V. Ng and C. Gardent. Improving machine learning approaches to coreference resolution. In *ACL*, pages 104–111, 2002.

[10] F. Niu, C. Zhang, C. Ré, and J. W. Shavlik. Felix: Scaling inference for markov logic with an operator-based approach. *CoRR*, 2011.

[11] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 2012.

[12] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* 2003.

[13] M. L. Wick, A. Culotta, K. Rohanimanesh, and A. McCallum. An entity based model for coreference resolution. In *SDM*, pages 365–376, 2009.

[14] S. Žitnik, L. Šubelj, D. Lavbič, O. Vasilecas, and M. Bajec. Contextual data matching and merging using semantics, trust and ontologies. *Informatica - in review*, 2012.