Structured Prediction for Joint Class Cardinality and Entity Property Inference in Model-Complete Text Comprehension

Hendrik ter Horst and **Philipp Cimiano**
Semantic Computing Group
Bielefeld University, Germany
{hterhors,cimiano}@techfak.uni-bielefeld.de

Abstract

Model-complete text comprehension aims at interpreting a natural language text with respect to a semantic domain model describing the classes and their properties relevant for the domain in question. Solving this task can be approached as a structured prediction problem, consisting in inferring the most probable instance of the semantic model given the text. In this work, we focus on the challenging subproblem of *cardinality prediction* that consists in predicting the number of distinct individuals of each class in the semantic model. We show that cardinality prediction can successfully be approached by modeling the overall task as a joint inference problem, predicting the number of individuals of certain classes while at the same time extracting their properties. We approach this task with probabilistic graphical models computing the maximum-a-posteriori instance of the semantic model. Our main contribution lies on the empirical investigation and analysis of different approximative inference strategies based on Gibbs sampling. We present and evaluate our models on the task of extracting key parameters from scientific full text articles describing pre-clinical studies in the domain of spinal cord injury.

1 Introduction

While there has been significant progress on information extraction tasks with a comparably low level of structural complexity such as entity recognition (Goulart et al., 2011; Nadeau and Sekine, 2007), relation extraction (Zhou et al., 2014; Kumar, 2017), and co-reference resolution (Soon et al., 2001; Ferracane et al., 2016), there is not much progress on capturing the comprehensive meaning of a text with respect to a given semantic model in terms of a given vocabulary of classes and properties. We refer to this task as *model-complete text comprehension* (MCTC) which requires to put all the above mentioned classical NLP-tasks into a larger context. The goal of MCTC is to capture all the information in the text that is expressible with respect to the semantic model, while ignoring those meaning aspects which are not. This can be framed as a structured prediction problem consisting in inferring the most plausible instance of the semantic model.

One challenging problem in MCTC lies in the prediction of the correct number of individuals for each class, hereinafter referred to as *cardinality prediction*, that is answering the question(s): "*How many (and which) individuals of a class are mentioned in the text?*". In essence, this can be approached by grouping mentions of known real-world entities into equivalence classes, which has widely been addressed under the heading of *co-reference resolution* (He, 2007; Singh et al., 2013). However, in many problem domains, we need to identify equivalence classes of entities that are priorily unknown (in terms of not referring to a specific real-world entity). Thus, explicit mentions in text such as naming variations etc. can not be directly mapped to a set of existing entities. To the contrary, such entities are only distinguishable on the basis of their describing properties. Take the case of scientific publications concerning pre-clinical studies containing a variable number of experimental groups each of which is described by an injury model, an animal species, treatments etc. Here, mentions of experimental groups do not refer to existing real-world entities and they need to be inferred/grouped on the basis of their identifying properties that are mentioned in the text. We refer to the prediction of how many *distinct* individuals[1] of a particular class are (indirectly) mentioned in a text as cardinality prediction and solve it jointly

[1] We refer to mentions of entities in a text as *entities* and to the denotation of such entities in a given model of the text as *individuals*

Proceedings of 4th Workshop on Structured Prediction for NLP, pages 22–32
November 20, 2020.

with the prediction of the properties of each individual. We model this joint task as the task of predicting a (logical) model of the text, which involves making choices as to which individuals exist for each class.

Towards capturing the dependence of class cardinalities and properties, we propose a joint inference approach that infers equivalence classes of entities in a text while at the same time predicting the properties of each equivalence class. We model this task as a statistical inference problem, relying on a factorized posterior conditional distribution $p(\vec{y} \mid \vec{x})$ as implemented in CRFs to approximate the true distribution over possible instantiations $\vec{y} \in Y$ of the semantic model given a text $\vec{x}$. Applying maximum-a-posteriori inference, we infer the most likely instance of the model that captures the whole meaning of the text as expressible by the semantic model. This includes the determination of the number of distinct equivalence classes (thus solving cardinality prediction) as well as predicting the properties for each equivalence class. Our approach is evaluated on text comprehension of research articles describing pre-clinical studies in the domain of spinal cord injury. Capturing correct key-parameters of the study protocol can be modeled as an MCTC problem as it requires a comprehensive understanding of the text rather than extracting single binary relations only. In this domain, we focus in particular on the extraction of experimental groups and their properties as described in Section 4.1. The data set[2] and the source code [3] are public available.

In this work, we answer the following research questions:

1) What is the advantage of jointly predicting the cardinality of classes and their properties over an isolated approach and how much does the prediction of the cardinality profit from the joint modelling?

2) What approximative inference strategies work best on this complex inference problem? We examine i) a vanilla Gibbs-based inference strategy ii) an inference strategy that is seeded with cardinality values based on a preceding clustering step., and iii) a parallel multi-chain inference strategy in which one chain is constructed for each potential cardinality value.

2 Related Work

There are a number of traditional natural language processing tasks related to *model-complete text comprehension*. In this section, we briefly discuss each task and provide some pointers to systems addressing the corresponding task, focusing on the bio-medical domain.

Entity Recognition and Linking (NER+L) describes the task of finding entity mentions in a text and linking them to unique concepts in some knowledge base. The task originated in the context of information extraction, consisting of identifying persons, company names etc. (Nadeau and Sekine, 2007) but has also received prominent attention in the biomedical field focusing on entities such as genes, diseases, treatments, etc. (Goulart et al., 2011). NER+L is an important preliminary step in many downstream applications as it identifies core informational units that are needed for more complex analysis levels including relation extraction, slot filling, and MCTC.

Relation Extraction (RE) describes the task of detecting relations between entities mentioned in a text (Giuliano et al., 2007). While many models rely on a pipeline architecture predicting entities first and then predicting relations, more recent works model both tasks jointly (Luo et al., 2015). Although there has been notable progress on RE in the last years, the task has been typically restricted to extracting binary relations within single sentence boundaries only (Zhou et al., 2014). With our work, we strive to go beyond such simplifications towards document-level text interpretation with respect to a more complex model.

Co-reference resolution (CRR) describes originally the task of finding nouns and pronouns that refer to the same underlying entity (Soon et al., 2001). When applying CRR to the medical field, the task shifts towards the resolution of mentions of diseases, tests, compounds, groups, treatments, etc. (He, 2007). Cardinality prediction in isolation can be modeled as a CRR problem, where the number of distinct non co-referring entities need to be found. With regard to the goal of comprehensive text understanding, classical co-reference resolution is clearly not enough, as also the properties of each entity need to be extracted. While Singh et al. (Singh et al., 2013) have attempted

[2] `http://psink.techfak.uni-bielefeld.de/spnlp-2020/mctc-spnlp2020.zip`

[3] `https://github.com/ag-sc/SCIOExtraction`

to model the tasks of entity recognition, relation extraction and co-reference resolution jointly, in their approach the interaction between relation extraction and co-reference resolution is not modelled directly, only via entity tags. In our approach we model the joint interaction between inducing equivalence classes (resolving co-references) while extracting the properties of entities/individuals as a basis to inform the decision about whether two individuals are the same (thus co-refer) given their properties. Durret et al. (Durrett et al., 2013) propose a global inference entity-level modeling for classical co-reference resolution based on a rich factor graph. In the unrolled factor graph, each factor refers to one entity property defined on a semantic or syntactic linguistic basis. In contrast to this work where properties of an individual/entity are pre-defined by the semantic model. Thus, our focus lies in their joint exploration while learning their interplay during inference in order to decide whether the properties belong to the same individual or not. Haghighi et al. (Haghighi and Klein, 2010) propose an unsupervised generative model incorporating several linguistic properties of the entity and its mention. In contrast, our work does not rely on entities that are explicitly mentioned in text. Instead, our model follows the schema of a semantic model to reason about the existence of individuals that can be inferred from the text and groups these individuals into groups by way of inferring the properties of these individuals.

The task of *slot-filling* (SF) was first introduced in the Message Understanding Conference (Grishman and Sundheim, 1996). It is concerned with predicting an entity-centric structure having a set of relations to other entities as it can be found e.g. in ontology-based information extraction (Sanchez-Cisneros and Aparicio Gali, 2013; Buitelaar et al., 2006) or extracting info-boxes from Wikipedia articles (Lange et al., 2010). Contrary to MCTC, classical slot-filling requires the prediction of a single structure per document only, which heavily reduces relational complexity and does not include nested individuals. There are many approaches to SF ranging from relying on distant supervision as described by Surdeanu et al. (Surdeanu et al., 2010) to, more recently, neural approaches as described by Zhang et al. (Zhang et al., 2017). Finally, SF can be seen as an upstream process for (cold-start) knowledge base population as described by ter Horst et al. (ter Horst et al., 2018).

Our work is highly related to information extraction systems in the (bio-) medical field. When it comes e.g. to the prediction of key parameters of clinical studies, most work focuses on the extraction of PICO-concepts: Patient/Problem (P), Intervention (I), Comparison (C) and Outcome (O). Summerscales et al. (Summerscales et al., 2009) have applied conditional random fields to extract key parameters from abstracts of clinical studies including treatments, experimental groups, and outcomes. Contrary to our approach, the task is defined as an NER+L problem, not aiming at capturing the semantic relations and concepts. Trenta et al. (Trenta et al., 2015) have proposed to rely on a maximum entropy classifier jointly extracting fine grained PICO elements from abstracts. Brujin et al. (De Bruijn et al., 2008) combined an SVM-based text classifier with regular expressions to extract PICO elements. Further, Ferracane et al. (Ferracane et al., 2016) aim to leverage co-reference resolution to identify experimental groups (patients) from medical abstracts. However, none of these works aims at deeper extraction of arms/experimental groups and their properties. In general, most approaches in the literature focus on sentence extraction and classification only (Mayer et al., 2018; Zhao et al., 2012; Wallace et al., 2016) rather than on predicting a semantic structure.

3 Method

Structured prediction describes a variety of tasks with the goal of predicting a pre-defined target structure that is extracted from an unstructured input text (Smith, 2011). We formulate the MCTC problem as a structured prediction task, where the structure to be predicted is an instance of the semantic model capturing the meaning of a text. This involves the task of predicting the number of individuals of each class (cardinality prediction) as well as predicting the values of the key properties of each individual. Our proposed method relies on probabilistic graphical models i.e. conditional random fields (CRFs; (Lafferty et al., 2001; Sutton et al., 2012)) as their application is well established in many structured prediction tasks in the context of NLP.

Encoding Semantic Models: An instance of the semantic model is encoded as a nested vector $\vec{y}$ containing as many elements as there are classes and properties in the model. Thus, given a set of classes $\{C_1, \ldots, C_n\}$ and a set of properties $\mathcal{P} =$

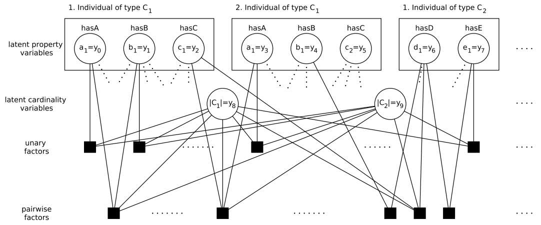

Figure 1: Schematized factor graph unrolled over the previously shown example. We introduce unary property factors connected to a single property of a single individual and pairwise property factors, connected to two properties of one or two individuals. Both factors are additionally connected to the cardinality variables jointly modelling the properties and cardinalities. For clarity, we omit the observed variables in this example.

$\{P_1, \ldots, P_m\}$, $\vec{y}$ can be written as $\{\vec{v}_{C_1}, ..., \vec{v}_{C_n}\}$ where each $\vec{v}_{C_i}$ has the form $[|C_i|, \vec{I^i_1}, \ldots, \vec{I^i_m}]$. $|C_i|$ represents the cardinality of class C_i, i.e. the number of individuals of class C_i mentioned in the text. $I^i_j \subseteq \mathcal{P}$ is a vector describing an individual of class C_i in terms of its properties.

Example: Consider a semantic model consisting of two classes C_1 and C_2 where individuals of class C_1 have properties $hasA, hasB, hasC$, and individuals of class C_2 have properties $hasD, hasE$. One specific instance of the semantic model would be represented as: $[[2, [a_1, b_1, c_1], [a_1, b_1, c_2]], [1, [d_1, e_2]]]$. The first component of the first tuple shows that there are two individuals of class C_1. The first individual has the property values a_1, b_1, c_1 for properties $hasA, hasB, hasC$, respectively. The second individual of class C_1 has property values a_1, b_1, c_2 for the above mentioned properties. The second tuple shows that there is one individual of class C_2 which has property values d_1, e_2 for properties $hasD, hasE$, respectively.

3.1 CRF-based Modelling

Let Y be the set of all possible (nested) vectors over a given vocabulary of classes and properties as exemplified above. Intuitively, this is the set of all possible instantiations of the semantic model. With $\vec{x}$ being the set of observed input variables corresponding to the list of tokens of the input text, the conditional probability of a specific instance of the semantic model $\vec{y} \in Y$ is $p(\vec{y}|\vec{x};\theta)$, with θ being a learned model parameter vector. The best value assignment to the set of target variables, denoted as $\hat{\vec{y}}$, is found by maximum a-posteriori (MAP) inference as shown in Equation (1):

$$\hat{\vec{y}} = \underset{\vec{y} \in Y}{\operatorname{argmax}}\, p(\vec{y}|\vec{x};\theta) \tag{1}$$

As inference in high dimensional vector spaces is often intractable, conditional random fields decompose the joint probability into individual factors. The set of factors and their operating scope is defined by a factor graph (Kschischang et al., 2001; Koller and Friedman, 2009). A factor graph is a bipartite undirected graph $\mathcal{G} = (V, F)$ consisting of a set of factors F and a set of variables V defined as the union of the observed input and the target output variables $V = \vec{y} \cup \vec{x}$. A factor $\Psi \in F$ is a non-negative real-valued exponential function $\Psi : V \rightarrow \mathbb{R}_{\geq 0}$ that computes a scalar score based on a subset $\omega \subseteq V$ of random variables defining its operating scope $\Psi(\omega) = \exp(\langle f(\omega), \theta_\Psi \rangle)$, with $f(\cdot)$ representing a feature vector based on a set of indicator functions, and θ_Ψ referring to the set of model weights that are shared between factors of the same type.

In our approach, it is crucial to capture dependencies between multiple target variables, in particular between the variables representing the cardinalities of classes and variables representing the individuals' properties. For this reason, we introduce factors that model the interaction between all pairs of property variables while having access to the cardinalities. We schematize our factor graph in Figure 1, unrolled over the previously given example. Let $\vec{C}$ denote the vector of the cardinalities of all classes and $|C_i| \in \vec{C}$ the cardinality of class C_i. The decomposition of the conditional probability $p(\vec{y} \mid \vec{x};\theta)$ can be written as shown in Equation (2):

$$\frac{1}{Z} \prod_{y_i \in \vec{y}} \Big[\Psi'(\vec{C}, y_i, \vec{x}) \prod_{y_j \in \vec{y} \setminus \{y_i\}} \Psi''(\vec{C}, y_i, y_j, \vec{x}) \Big] \tag{2}$$

where Z denotes the partition function and $\Psi'(\cdot)$, $\Psi''(\cdot)$ denote factors defined for single and pairs of output variables while having access to the cardinalities $\vec{C}$.

The unrolling of factors over the input is performed using imperatively defined factor graphs as proposed by McCallum et al. (McCallum et al., 2009). For approximative inference of the posterior distribution, we rely on the state-based Markov Chain Monte Carlo sampling paradigm. Proposal states are computed and sampled via Gibbs sampling (Casella and George, 1992). While training, the model parameters θ are updated with SampleRank (Wick et al., 2009) that is computing parameter update gradients based on an objective comparison of two states, usually between the current state and the selected successor state (cf. next sections for proposed variations). In our approach the objective is to maximize the F_1 score to the ground truth.

3.2 State-based Inference Strategies

In the following, we propose our inference strategies to MCTC with a focus towards cardinality prediction. In state-based inference, a state s^t is defined as one specific variable assignment to the target structure $\vec{y}$ at a specific time point t. While inference proceeds, in each step a set of proposal states $\mathcal{S}^{t+1}$ is computed based on a list of predefined atomic change rules that are applied to the current state, e.g. changing cardinalities of classes or the properties of individuals. The successor state $s^{t+1} \in \mathcal{S}^{t+1}$ is sampled from the generated set of proposal states.

Vanilla Inference: The *vanilla inference* is based on traditional Gibbs sampling. The inference procedure is initialized with one empty state s^0 that is $\vec{y} = \emptyset$ (no values are assigned) and iteratively updated with atomic change rules. Modifying the cardinality for a class C_i is defined as either deleting an existing individual of index j ($\vec{y} \leftarrow \vec{y} \setminus \vec{I}^i_j$; $|C_i| \leftarrow (|C_i| - 1)$) or adding a new individual with leading index $|C_i|$ ($\vec{y} \leftarrow \vec{y} \cup \vec{I}^{\,i}_{|C_i|}$; $|C_i| \leftarrow (|C_i| + 1)$). On the level of individuals, an atomic change is defined as deleting, adding, or changing a property value. The inference procedure terminates if the model parameter update converges. The final state represents the most likely instance of the semantic model.

Cardinality Seeded$^+$ Inference: In the *seeded$^+$ inference* (c.f. Figure 2), the first state s^0 is initialized with an a priori predicted cardinality value λ_{C_i} for each class C_i, which is re-sampled as inference proceeds. For this, the system relies on the same atomic change and termination rules as defined for the vanilla inference.

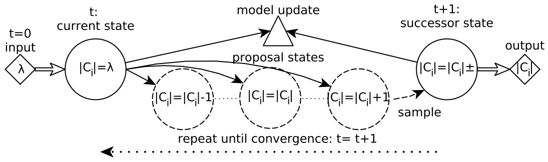

Figure 2: Schematized *seeded$^+$ inference*. The input is the seed parameter λ which is used to initialize the cardinality of the first state. Within the proposal states, λ can be altered. In each time step the model is updated based on the current state and successor state.

Parallel Multi Chain Inference: The *parallel multi chain inference* procedure (c.f. Figure 3) is initialized with n independent Markov chains $S^0 = [s^0_1, s^0_2, \ldots, s^0_n]$ that are explored in parallel but independently from each other. Each state $s^0_i \in S^0$ is initialized with a fixed number of individuals for each class type ranging between a pre-defined minimum α_{C_i} and maximum β_{C_i}. Contrary to the previous inference strategies, the cardinalities in each chain are not sampled over but remain fixed. Only the property values of the individuals are sampled. The parallel sampling is independent in the sense that for each chain the computation of the set of proposal states and the selection of the successor states is independent of the other chains. The model parameters θ however are shared throughout all chains and are thus updated n times every time step; once for each pair of current–successor state. This inference procedure terminates if all chains converge. The final output is selected based on the highest model probability among the final states of all chains.

Parallel Multi Chain Inference$^+$: The *parallel multi chain inference with cross-chain model updates* strategy builds on the previous inference strategy in that it includes parallel inference chains with fixed cardinality but integrates cross-chain model update operations after each time step (bold triangle in Figure 3). That is, in addition to the n model updates, a set of state-pairs is computed by pairwise combining the selected successor states of the chains for cross-over model updates. This generates $\frac{n^2+n}{2}$ model parameter updates in each time step. The motivation for this cross-chain model updates is to force the model to learn to prefer the correct cardinality values.

3.3 Features

Factors are defined in terms of indicator functions that measure the compatibility of variable assignments to the output structure $\vec{y}$ given the input document $\vec{x}$. In the following, we explain four types of feature groups that we consider in our model. The proposed features are intuitively designed to capture document-level semantics and finally selected empirically based on an evaluation of a subset of the training data.

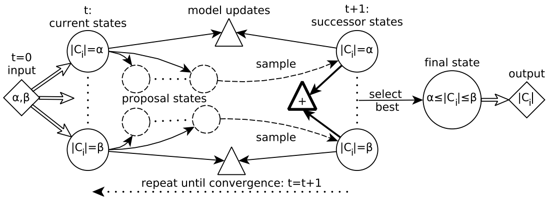

Figure 3: Schematized *parallel multi chain*$^+$ *inference*. For each value between the input α and β a Markov chain is instantiated. In each time step the model is updated between the current and successor states for each chain. In the advanced version, additional model update operations based on the successor states are added (bold triangle).

Document-level: Document-level features measure the compatibility of property assignments of individuals based on the textual content of the document represented as 3-grams. For this, triples are considered for plausibility, containing the property type, the entity type of the property value, and its textual representation as 3-grams. We further measure the compatibility of pairwise assignments of property values considering their sentential distance, assuming that values within the same property (in case of multi value properties) or individual (throughout multiple properties) are more likely to appear closer together rather than being spread across the document.

Document-structure: Document structure features rely on a heuristic segmentation of the document into the standard sections of a scientific article: *abstract*, *introduction*, *method*, *results*, *discussion*, *references*, and *unknown*. We compute features that capture 3-grams mentioned in specific sections of the article as indicators for the assignment of certain values to properties. By this, we can model that certain content is expected in certain sections and should override inconsistent information appearing in other sections.

Cardinality: Aiming at cardinality prediction, we measure the compatibility of cardinality values in dependence of other random variables in $\vec{y}$. For this, we make the choice of a cardinality dependent on n-grams appearing in the surface forms of property values.

In addition, we also consider features implementing a prior for the cardinalities of classes as well as for the number of different values for multi-value properties. By this, the model is able to learn a class/property-specific distribution of cardinality values. For example, assuming that the cardinality of a class $\hat{C}$ has a very high a priori likelihood for a specific value $\lambda_{\hat{C}}$ throughout the training data, this puts pressure on the model during inference to prefer model instances where there are $\lambda_{\hat{C}}$ individuals for the respective class, unless other features provide strong evidence for the contrary.

Within- and Across-Individual Coherence: Sometimes values of properties are shared across individuals within the same class. Thus, we measure the compatibility of value assignments across properties within one individual, but also how plausible it is that a certain value is shared across individuals.

4 Experiments

Model-complete text comprehension aims at the automatic instantiation of a semantic model based on information extracted from a natural language text. Such an instance contains information about individuals of equivalence classes, their cardinality and their properties. Thus the overall task of MCTC can be evaluated towards i) the correct prediction of the number of individuals, and ii) the prediction of properties for each individual. In the following, we describe our use case application, the experimental procedure and results.

4.1 Semantic Model and Data Set

We apply our approach to full text articles describing pre-clinical studies in the domain of spinal cord injury. Our semantic model is an excerpt of the Spinal Cord Injury Ontology (SCIO) (Brazda et al., 2017) centered on the key concept of an experimental group. An experimental group represent an animal model to which a certain injury and treatment is applied and is described by four key properties: i) *hasSpecies* specifying the species that the animal model belongs to, ii) *hasInjury* specifying the experimentally inflicted injury, ii) *hasTreatment* is the list of treatments that were applied, and iv) *hasName* is a list of naming variations for that animal group that are used throughout the document. Note that, in accordance with domain experts, only the first three properties are considered to be relevant to describe the experimental group semantically and thus are evaluated. However, the property *hasName* can be seen as an auxiliary property that is not necessary to understand the study but provides useful information, e.g. to detect co-references.

The data set contains full text articles that have been annotated by three domain experts using the

SANTO framework (Hartung et al., 2018). Annotations are available on the full level of relevant concepts of SCIO. Each document can be seen as a data point that is annotated with an instance of the semantic model containing a list of experimental groups and their properties. While annotations for the *hasName* property are linked to specific textual phrases in the document, all other properties are annotated in a distantly supervised fashion. In a preliminary step, we apply a named entity recognition heuristic based on automatically generated regular expressions to compute a set of document-based annotations for all classes and property values that exist in the semantic model. The names of groups are additionally extracted with a standard CRF using standard token-level features. The final data set contains 96 data points with an average length of approx. 273 sentences per document and a total number of 345 experimental groups ($\mu = 3.3,\ \sigma = 1.3,\ min = 2,\ max = 7$).

4.2 Inference Parameter Estimation

Our proposed inference strategies rely on a prior estimation of the number of individuals for initialization. As described in Section 3.2 the seeded inference strategy requires the seed variable λ. The parallel multi chain inference requires a range of cardinality values $0 \leq \alpha \leq \beta$. Details about their estimation are briefly described below.

Seed Prior Cardinality Estimation λ The cardinality seeded$^{(+)}$ inference procedure requires the estimation of the seed parameter λ for each class determining the number of individuals (experimental groups) the initial state begins with. λ is computed by relying on the k-Means algorithm by clustering group names based on textual features. The cluster quality of k-Means depends on two main parameters. First, the determination of the number of clusters, for which we rely on the residual sum of squares (RSS) algorithm with an empirically determined penalization factor for large number of clusters. Second, we rely on a function measuring the distance between two data points, i.e. between two group names. We compare three distance functions: i) Levenshtein distance with a k-Medoid implementation of k-Means, ii) cosine distance of the averaged sum of pre-trained Pubmed-based word embeddings induced with Word2vec (Mikolov et al., 2013), and iii) a random forest classifier (Liaw et al., 2002) with a correlation based feature selection (resulting in Smith-Waterman and 3-gram based Jaccard similarity as features).We evaluate the performances based on the F$_1$ score using a reference clustering as ground truth obtained from our annotated data set. We define a true positive as a group name that is in the correct cluster, a false positive if it is in a wrong cluster, and a false negative if it is missing in its respective cluster. The Levenshtein distance performed with F$_1$ $= 0.41$, the Word2Vec-based cosine distance performs slightly better with F$_1$ $= 0.45$, while the random forest classifier reaches a value of F$_1$ $= 0.56$. With the random forest outperforming both other models, we rely on this distance function in a k-Means clustering for estimating λ.

Parallel Multi Chain$^{(+)}$ Parameters α and β The parallel multi chain$^{(+)}$ inference strategies require the estimation of a minimum (α) and a maximum (β) number of individuals assuming that the correct cardinality lies between α and β. We estimate both parameters in dependence of the average cardinality of individuals in the training set. With μ being the average cardinality and σ its standard deviation, we set $\alpha = \mu - \sigma$ and $\beta = \mu + \sigma$.

4.3 Evaluation Setting

Our experiments follow a randomized cross validation regime as usual for experiments on relatively small data sets. We ran each experiment 10 times with a random split into 80% training data (76 in number) and 20% test data (20 in number). We provide evaluation results in terms of precision, recall, and F$_1$ macro averaged over all documents in three configurations:

Cardinality Prediction (CP): We compare the predicted cardinality p_c to the ground truth cardinality g_c where $tp = min(p_c, g_c), fp = max(0, (p_c - g_c))$, and $fn = max(0, (g_c - p_c))$

Property Prediction (PP): We compare the predicted property values to the ground truth property values where a true positive is a correctly assigned property value, a false positive is a wrongly assigned property value and a false negative as a missing property value of an individual.

Combined (Comb): We compute the harmonic mean between the cardinality and property prediction scores.

4.4 Experimental Results

Our experiments comprise the evaluation of four models each of which is based on one of the described joint inference strategies predicting cardinality and properties at the same time, as well as a pure cardinality prediction baseline ignoring property prediction. The joint inference models are: *RSS*: the seeded inference with a fixed cardinality, *RSS*$^+$: the seeded inference that allows further sampling of the cardinality as described in Section 3.2, *PAR*: the parallel multi chain inference, and *PAR*$^+$: the parallel multi chain inference with chain-cross over model updates. As cardinality baseline(s), we provide *Co-ref CRF*, a CRF based method for clustering group names without a joint prediction of properties, relying on linguistic features only, and *RSS*, the cardinality as predicted by the RSS based k-Means as reported in the RSS-model. Note that the cardinality in RSS is fixed and does not change during inference so that it can be seen as a baseline for predicting the cardinalities. The experimental results of those models are reported in Table 1. In Table 2, we compare the run time and the number of generated states for the four inference methods.

4.5 Discussion

We analyze the results with respect to three different aspects: i) performance of the cardinality prediction, ii) overall performance as measured by the combined harmonic mean, and iii) performance with respect to the run time and complexity.

Cardinality Prediction The performances of the cardinality prediction can be seen in the first row of Table 1. The CRF-based baseline already yields a very strong F_1-score of 0.79 which shows that cardinality prediction with linguistic features ignoring property prediction provides already decent results. The k-Means approach with an unsupervised RSS cluster estimation yields a cardinality F_1-score of 0.64, performing worse than the CRF baseline. When seeding our approximate inference approach with prior cardinality values (RSS$^+$), the F_1-score considerably improves by 19 %-points up to 0.83, even outperforming the CRF baseline. The cardinality prediction in PAR performs comparably strong with an F_1-score of 0.81. This score is further outperformed when integrating the cross-chain model update operation. PAR+ archives performs best in predicting the cardinalities with an F_1-score of 0.84.

Overall Score The performances of the overall prediction can be seen in the second to last rows in Table 1. With respect to the property prediction, RSS performs best with a score of 0.57, mainly due to the correct detection of TREATMENTS (0.67) and SPECIES (0.62). With a low cardinality performance, the overall score sums up to 0.63 in F_1. The strong increase in the performance of cardinality prediction in RSS$^+$, compared to the RSS model comes at the cost of an inferior property prediction quality. The combined score for RSS$^+$ however shows slightly better results with an F_1-score of 0.65. The property prediction in PAR shows similar results to the RSS$^+$ for INJURY, a slight decrease for TREATMENTS, and a huge increase (10% points) for SPECIES. The PAR model yields an overall score of 0.66. Activating cross-chain model updates (PAR$^+$), the property prediction shows a performance increase by 8% points for *hasTreatment* while for both other properties the value is similar to PAR. The PAR$^+$ model outperforms all other models in the overall score, but lacks 4%-points for property prediction in comparison to RSS. The results show that property prediction works best when fixing the number of individuals. With PAR$^+$ model working best for cardinality prediction, an interesting model combination could be to use the cardinality output of PAR$^+$ as initialization to RSS. This however, is left for future work.

Run Time Performance The run time as well as the number of states for each inference method is shown in Table 2. We report statistics on the average time in seconds (s) that is needed to process a document and depict the search space complexity by providing the average number of generated and evaluated states in thousands (k). All experiments ran on an Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz with 16 cores and 120 GB of available RAM. No GPU or further hardware acceleration was used. The table shows that RSS has the lowest complexity in terms of state generation, which is due to the fixed cardinality and in consequence a significantly reduced search space. In RSS$^+$, we notice a huge increase in the number of generated states by a factor of around 7. At the same time, we observe that the run time factor rises only by a factor of 2.2 in training and 2.8 in test. It is noticeable that the number of generated states and the run time at test time decreases from PAR to PAR$^+$ which is probably due to a faster model convergence, however training run time increases.

Approach	Co-ref CRF			RSS			RSS$^+$			PAR			PAR$^+$		
	F$_1$	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**
CP	0.79	0.99	0.65	0.64	0.48	0.97	0.83	0.89	0.78	0.81	0.70	0.96	**0.84**	0.75	0.96
Comb				0.63	0.53	0.77	0.65	0.68	0.63	0.66	0.60	0.73	**0.69**	0.64	0.75
PP				**0.57**	0.58	0.57	0.48	0.47	0.48	0.50	0.50	0.49	0.53	0.53	0.53
→Injury				0.35	0.35	0.35	0.46	0.46	0.46	0.47	0.47	0.47	**0.48**	0.48	0.48
→Species				**0.62**	0.62	0.62	0.50	0.50	0.50	0.60	0.60	0.60	0.61	0.61	0.61
→Treatments				**0.67**	0.68	0.66	0.46	0.45	0.48	0.42	0.44	0.41	0.50	0.51	0.49

Table 1: Results of the cardinality baseline(s) and of the inference strategies for joint cardinality and property prediction. We provide macro-F$_1$, precision, and recall averaged over 10 runs with random 80/20 splits.

Approach	RSS	RSS$^+$	PAR	PAR$^+$
Avg. # states (k)	**46**	324	150	119
Avg. train time (s)	**28.11**	63.87	54.71	60.73
Avg. test time (s)	**8.95**	25.21	38.67	32.87

Table 2: Run time and complexity statistics of the inference strategies. We provide the average number of evaluated states in thousands (k), averaged training and test time per document in seconds (s).

5 Conclusion

We have proposed an approach to the task of *model-complete text comprehension* (MCTC) that relies on a learned model of the posterior distribution of instances of a semantic model given a text to infer the most likely instance of a semantic model that captures the meaning of the text best. We have relied on CRFs to model the conditional distribution in a factorized way and empirically investigated the impact of different approximate inferences strategies on our problem. Our experiments on the task of predicting the structure of experimental groups from scientific full text articles describing pre-clinical studies in the field of spinal cord injury show that modeling the MCTC task as a joint inference problem, extracting the cardinality in combination with predicting the properties of the individuals, outperforms a number of reasonable baselines predicting the cardinality alone. In future work, we intend to investigate combinations of our inference strategies, relying on the result state produced by our PAR$^+$ inference strategy to seed the RSS inference method to re-sample the property values, expecting to see an overall gain in both cardinality prediction and entity property prediction over both inference strategies.

Acknowledgements

This work has been funded by the Federal Ministry of Education and Research (BMBF, Germany) in the PSINK project (project number 031L0028A).

References

Nicole Brazda, Hendrik ter Horst, Matthias Hartung, Cord Wiljes, Veronica Estrada, Roman Klinger, Wolfgang Kuchinke, Hans Werner Müller, and Philipp Cimiano. 2017. Scio: an ontology to support the formalization of pre-clinical spinal cord injury experiments. In *Proc. of the 3rd Joint Ontology Workshops (JOWO): Ontologies and Data in the Life Sciences*, volume 2050.

Paul Buitelaar, Philipp Cimiano, Stefania Racioppa, and Melanie Siegel. 2006. Ontology-based information extraction with soba. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*.

George Casella and Edward I George. 1992. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Berry De Bruijn, Simona Carini, Svetlana Kiritchenko, Joel Martin, and Ida Sim. 2008. Automated information extraction of key trial design elements from clinical trial publications. In *Proc. of the AMIA Annual Symposium*, volume 2008, page 141. American Medical Informatics Association.

Greg Durrett, David Hall, and Dan Klein. 2013. Decentralized entity-level modeling for coreference resolution. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 114–124.

Elisa Ferracane, Iain Marshall, Byron C Wallace, and Katrin Erk. 2016. Leveraging coreference to identify arms in medical abstracts: An experimental study. In *Proc. of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 86–95.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2007. Relation extraction and the influence of automatic named-entity recognition. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(1):1–26.

Rodrigo Rafael Villarreal Goulart, Vera Lúcia Strube de Lima, and Clarissa Castellã Xavier. 2011. A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society*, 17(2):103–116.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *Proc. of the 16th International Conference on Computational Linguistics (COLING)*.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393.

Matthias Hartung, Hendrik ter Horst, Frank Grimm, Tim Diekmann, Roman Klinger, and Philipp Cimiano. 2018. Santo: a web-based annotation tool for ontology-driven slot filling. In *Proceedings of ACL 2018, System Demonstrations*, pages 68–73.

Tian Ye He. 2007. *Coreference resolution on entities and events for hospital discharge summaries*. Ph.D. thesis, Massachusetts Institute of Technology.

Hendrik ter Horst, Matthias Hartung, and Philipp Cimiano. 2018. Cold-start knowledge base population using ontology-based information extraction with conditional random fields. In *Proc. of the Reasoning Web International Summer School (RW)*, pages 78–109. Springer.

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models. Principles and Techniques*. MIT Press.

Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. 2001. Factor Graphs and Sum Product Algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

Shantanu Kumar. 2017. A survey of deep learning methods for relation extraction. *CoRR*, abs/1705.03645.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields. Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 282–289.

Dustin Lange, Christoph Böhm, and Felix Naumann. 2010. Extracting structured information from wikipedia articles to populate infoboxes. In *Proc. of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1661–1664.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 879–888.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2018. Evidence type classification in randomized controlled trials. In *Proc. of the 5th Workshop on Argument Mining*, pages 29–34. Association for Computational Linguistics.

Andrew McCallum, Karl Schultz, and Sameer Singh. 2009. Factorie: Probabilistic programming via imperatively defined factor graphs. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 1249–1257.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Daniel Sanchez-Cisneros and Fernando Aparicio Gali. 2013. UEM-UC3M: An ontology-based named entity recognition system for biomedical texts. In *Proc. of the Seventh International Workshop on Semantic Evaluation (SemEval)*, pages 622–627. Association for Computational Linguistics.

Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. 2013. Joint inference of entities, relations, and coreference. In *Proc. of the 2013 workshop on Automated knowledge base construction (AKBC)*, pages 1–6.

Noah A. Smith. 2011. *Linguistic Structure Prediction*. Morgan and Claypool.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Rodney Summerscales, Shlomo Argamon, Jordan Hupert, and Alan Schwartz. 2009. Identifying treatments, groups, and outcomes in medical abstracts. In *Proc. of the Sixth Midwest Computational Linguistics Colloquium (MCLC)*. Indiana University.

Mihai Surdeanu, David McClosky, Julie Tibshirani, John Bauer, Angel X Chang, Valentin I Spitkovsky, and Christopher D Manning. 2010. A simple distant supervision approach for the tac-kbp slot filling task.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Antonio Trenta, Anthony Hunter, and Sebastian Riedel. 2015. Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints. *CoRR*, abs/1509.05209.

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, 17(1):4572–4596.

M. Wick, K. Rohanimanesh, A. Culotta, and A. McCallum. 2009. SampleRank. Learning Preferences from Atomic Gradients. In *Proc. of the NIPS Workshop on Advances in Ranking*, pages 1–5.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 35–45.

Jin Zhao, Praveen Bysani, and Min-Yen Kan. 2012. Exploiting classification correlations for the extraction of evidence-based practice information. In *Proc. of the AMIA Annual Symposium*, volume 2012, page 1070. American Medical Informatics Association.

Deyu Zhou, Dayou Zhong, and Yulan He. 2014. Biomedical relation extraction: from binary to complex. *Computational and Mathematical Methods in Medicine*, 2014.