Grammar and Dictionary based Named-entity Linking for Knowledge Extraction of Evidence-based Dietary Recommendations

Tome Eftimov[1,2], Barbara Koroušič Seljak[1] and Peter Korošec[1,3]
[1]*Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia*
[2]*Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia*
[3]*Faculty of Mathematics, Natural Sciences and Information Technologies, Glagoljška 8, 6000 Koper, Slovenia*

Keywords: Named-entity Linking, Knowledge Extraction, Dietary Recommendations, Computational Linguistics, Public Health.

Abstract: In order to help people to follow the new knowledge about healthy diet that comes rapidly each day with the new published scientific reports, a grammar and dictionary based named-entity linking method is presented that can be used for knowledge extraction of evidence-based dietary recommendations. The method consists of two phases. The first one is a mix of entity detection and determination of a set of candidates for each entity, and the second one is a candidate selection. We evaluate our method using a corpus from dietary recommendations presented in one sentence provided by the World Health Organization and the U.S. National Library of Medicine. The corpus consists of 50 dietary recommendations and 10 sentences that are not related with dietary recommendations. For 47 out of 50 dietary recommendations the proposed method extract all the useful knowledge, and for remaining 3 only the information for one entity is missing. Due to the 10 sentences that are not dietary recommendation the method does not extract any entities, as expected.

1 INTRODUCTION

Food-based dietary guidelines (FBDGs) are simple advices on healthy eating, aimed at the general public (Vorster et al., 2001). They give an indication of what a person (an individual member of the general public) should be eating in terms of foods, and provide a basic framework to apply when making healthy dietary choices and planning meals. The main goal of FBDGs is to improve public health and well-being.

The European Food Safety Agency (EFSA) (EFSA, 2016) is an example of the authority, which provides dietary reference values (DRVs) as a complete set of nutrient recommendations and quantitative reference values for nutritional intakes, such as population reference intakes, the average requirement, adequate intake level, and the lower threshold intake. DRVs form a basis for establishing FBDGs, which translate nutritional recommendations into messages about foods and diet.

Most countries have established their own national DRVs and FBDGs that consider beside international recommendations and guidelines also local conditions and national/ethnical eating culture and habits, and are reviewed and updated from time to time.

As today the amount of information is massive and is quickly increasing, computer-based tools for systematic knowledge identification, extraction and exploration are welcome to support human experts in decision-making about appropriate nutritional care for specific disease states or conditions in typical settings.

In this paper we present a method that is a grammar and dictionary based named-entity linking and could be used for knowledge extraction of evidence-based dietary recommendations. In Section II we review the appropriate related work. Section III describes the problem in depth. Section IV describes the proposed method in general. In section V the evaluation and results of the method are presented, and in Section VI we conclude the paper by discussing the importance of this method and our plans for future work.

2 RELATED WORK

Entity-linking is a natural language processing task that is used for identifying text strings that refer

Eftimov, T., Seljak, B. and Korošec, P.
Grammar and Dictionary based Named-entity Linking for Knowledge Extraction of Evidence-based Dietary Recommendations.
DOI: 10.5220/0006032401500157
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 150-157
ISBN: 978-989-758-203-5
Copyright © 2016 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

to a particular item in some reference knowledge base (Mihalcea and Csomai, 2007; Han et al., 2011; Hachey et al., 2013; Blanco et al., 2015). Usually it happens in three different phases. The first phase is a entity detection in which the linker identifies which parts of the text are likely to refer to an entity. In the second phase candidates nodes for each entity are determined. The third phase is a candidate selection in which the true candidates for each entity are selected from the set of its candidates.

3 PROBLEM DEFINITION

The problem we are interested is knowledge extraction of evidence based dietary recommendations that are represented as a simple sentence. Having the information represented as text, we need to catch the information related to food, chemical components (nutrients), the quantity-unit pair that is recommended, and also the life stage group for which those recommendations are given. The first question is how to select the parts of the text (phrases) that will be the candidates for some entity that we are interested. For this we need to find a good way of tokenization, such that each phrase can be a candidate for an entity and a phrase must not contain information about more entities. For example if we have the dietary recommendation "*Some breakfast cereals contain 150 to 300 mg of sodium before milk is added*", we do not like to obtain token as "150 to 300 mg of sodium", because in this case we will have information about the quantity-unit pair and the chemical component in the same phrase. For us a better choice is "150 to 300 mg" and "sodium" to be separate phrases. Then we need to find good knowledge base to which each phrase will be linked in order to find the set of candidates for each entity. The most important step is to find the action from the recommendation (for example, "contain", "consist", "should further reduce", among others) in order to know what we need to do with the recommendation reported. The last phase is to find how we can select and extract only the true candidates for each entity from the set of its candidates.

4 GRAMMAR AND DICTIONARY BASED NAMED-ENTITY LINKING

Our proposed method is a grammar and dictionary based named-entity linking that can be used for knowledge extraction of evidence-based dietary recommendations. It consists of two phases. The first phase is a mix of entity detection and determination of a set of candidates for each entity, while the second phase is a candidate selection.

Let Φ be a simple sentence that contains information about dietary recommendation. We start by introducing the word-level tokenization on Φ. The result is a $n \times 1$ vector, *Words*, whose elements are the words from Φ, and n is the number of words obtained after the tokenization. Then we continue with POS (part-of-speech) tagging (A.Voutilainen, 2003; Schmid, 1994; Tian and Lo, 2015) that is a process of classifying words into their parts of speech or lexical categories, such as nouns (*NN*), verbs (*VB*), adverbs (*RB*), adjectives (*JJ*), among others, and information about subtagging, such as verbs' tenses, plural or singular noun, etc. The result is a $n \times 1$ vector, POS_{Tags}, which is a collection of POS tags for Φ. More details about the POS tags can be found in the Penn Treebank Project (Santorini, 1990; Marcus et al., 1993). After processing the sentence on word level, we continue with chunking, which segments and labels multitoken sequences called chunks that can be noun phrase (*NP*), verbal phrase (*VP*), adjective phrase (*ADJP*), prepositional phrase (*PP*) among others (Chowdhury, 2003). The result is a $n \times 1$ vector, *Chunks*, whose elements are chunk tokens tagged in IOB format (Inside(I-), Outside(O), Beginning(B-)). The B- prefix in a chunk token indicates that the chunk token is a beginning of the chunk, the I- prefix before a chunk token indicates that the chunk token is inside the chunk, and the O tag indicates that a chunk token belongs to no chunk.

The next step is to define a $n \times m$ matrix, X_{Chunks}, where m is the number of chunk tokens from the *Chunks* vector that begin with the prefix B- or O. The elements of the matrix X_{Chunks} are defined with the Equation 1, so if the word belongs to the chunk we have 1, and 0 otherwise.

$$X_{\text{Chunks}}[i,j] = \begin{cases} 1, & \text{if } Words[i] \in Chunks[j] \\ 0, & otherwise \end{cases}, \quad (1)$$

where $i = 1,..,n$ and $j = 1,..,m$.

Let k be the number of entities we are interested in. In order to define the set of candidates for each entity, we try to link each chunk with the information from additional dictionaries related to the domains of the entities, $Dictionary^l$, $l = 1,..,k$. These dictionaries can be vocabularies, ontologies, or semantic annotators. Then a $n \times k$ matrix, X_{Entities}, is defined as

$$X_{\text{Entities}}[i,l] = \begin{cases} 1, & \text{if } Words[i] \in Dictionary^l \\ 0, & otherwise \end{cases}. \quad (2)$$

After obtaining the matrices X_{Chunks} and X_{Entities}, a $m \times k$ matrix, $X_{\text{Candidates}}$, is defined as

$$X_{\text{Candidates}} = X_{\text{Chunks}}^{T} \cdot X_{\text{Entities}}. \quad (3)$$

The rows of the matrix $X_{\text{Candidates}}$ correspond to the chunks and the columns are the dictionaries we included as reference knowledge bases. For example if the element $X_{\text{Candidates}}[i,l] \geq 1$ then the i-th chunk is a candidate solution for the l-th entity.

To further improve the quality of the solution three additional chunkings are introduced. In the first chunking, trigrams of successive chunks $(Chunk_i, Chunk_{i+1}, Chunk_{i+2})$ are analyzed and merged into one new noun chunk if the trigram is composed as $(B-NP, B-PP, B-NP)$, expect in the cases when the two noun chunks correspond to candidates of different entities because merging them can cause losing information about one entity described by one of the noun chunks. In order not to lose this information, in Table 1 a boolean function together with the boolean variables A and B, that can be different entities in our case, is presented. Further, a Karnaugh map presented in Figure 1, also known as K-map (Nelson, 1955), is used to simplify boolean algebra expression when this chunking needs to be performed.

Table 1: The Boolean function for the first additional chunking.

A	B	f
0	0	1
0	1	1
1	0	1
1	1	0

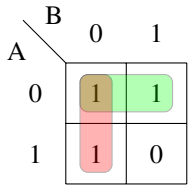

Figure 1: Karnaugh map of the Boolean function for the first additional chunking.

The boolean algebra expression or the boolean function when this chunking needs to be performed is obtained in simplified form as a sum of minterms as

$$f(A,B) = \overline{A} \vee \overline{B}. \quad (4)$$

Because the number of entities can be greater than 2, $k > 2$, the boolean algebra expression obtained with the Equation 4 needs to be defined for each variation of the pair of entities. The number of such obtained functions is determined using the formula of the variations without repetition $V_{r,w} = \frac{r!}{(r-w)!}$, where r is the number of different elements, in our case the number of different entities, k, and w is the size of variation or how many elements need to be selected from the set of r elements. In our case w is 2 because we are working with pair of entities. Then all the obtained functions are merged together with boolean *AND* conditions into one expression. This boolean algebra expression defines wheather the first chunking needs to be performed.

In the second additional chunking, trigrams of successive chunks $(Chunk_i, Chunk_{i+1}, Chunk_{i+2})$ are analyzed and merged into one new noun chunk if the trigram is composed as $(B-NP, B-VP, B-NP)$ and the first noun chunk has POS tag that is Wh-pronoun (Santorini, 1990), such as who, what, which, etc.

In the third additional chunking, bigrams of successive chunks $(Chunk_i, Chunk_{i+1})$ are merged into one new noun chunk if the bigram is composed as $(B-NP, B-NP)$, and only one of the noun chunks is labeled as entity of interest, or both of them have the same label.

After performing the three additional chunkings, the matrix X_{Chunks} needs to be recalculated because the number of chunks is different from the number obtained by the default chunking. At the end of the first phase of the method, the matrix $X_{\text{Candidates}}$ is recalculated and their columns correspond to the sets of candidates for each entity, respectively.

The next step of the named-entity linking is to select the true candidates from the set of candidates for each entity. For this purpose, the sentence is represented as graph, in which each chunk is connected only with its neighbours (the predecessor and the successor chunk). Then the start or initial node of the graph from where the search for all entities will start needs to be selected. The initial node of the graph is selected using a syntactic bracketing or tree parser (Taylor et al., 2003).

Each sentence Φ is represented by the parser as a tree having three children: a noun phrase (NP), a verbal phrase (VP) and a full stop (.). In addition, *Subject*, *Predicate*, and *Object* is the combination of three words that form any sentence (Rusu et al., 2007). The *Subject* is the person or a thing who or which carries out the action of the verb. The *Predicate* in a sentence is what tells about what a person or a thing does or did, or what happened to a person or to a thing. The *Object* is the person or a thing upon whom or upon which the action of the verb is carried out.

The initial node of the graph is the predicate of the sentence. The search for the predicate is performed in VP. The initial node can be found in the follow-

ing subtrees *VB* (verb, base form), *VBD* (verb, past tense), *VBG* (verb, present participle or gerund), *VBN* (verb, past participle), *VBP* (verb, present tense, not 3rd person singular), *VBZ* (verb, present tense, 3rd person singular) and *MD* (verb, modal). Further, from all candidates returned by searching for predicate, the initial node is the verb chunk that is closest to the root of the sentence (number of edges from the verb node to the root node) and it is located in verbal phrase that is closest to the root. The extracted predicate is stored in an entity called *Action*.

After the selection of the *Action* entity, all other entities of interest need to be selected. Because it can happened that no candidate is subject in the sentence, one additional entity called *Group* is added, into which the noun chunks that perform the action are stored. The *Group* entity is searched in the predecessor chunks from the *Action* entity that is selected. The searching starts from the *Action* entity and it goes back to the beginning of the sentence. The result are the successive noun chunks that can also be separated by punctuation. In order to know on which side of the *Action* entity the extracted entities are located and to catch the relations between them, one of the labels *S*, *P*, or *O* that indicate $(Subject, Predicate, Object)$ is added to each extracted entity. The *Action* entity has the label *P* because it is the predicate of the sentence. All entities that are predecessor chunks of the *Action* entity have label *S* or they are subjects in the sentence. The entities that are successor chunks of the *Action* entity have label *O* because they correspond to the objects in the sentence.

Two scenarios of the candidate selection exist.

In the first one if the *Action* entity is not selected, then all the candidate solutions from the $X_{\text{Candidates}}$ matrix are extracted.

In the second scenario, only one *Action* entity is returned. Then for each entity using the set of its candidates, the candidate or the chunk that is closest to the *Action* entity is selected, according to the number of edges between the candidate and the *Action* entity in the graph. If the set of candidates consists of more candidates for the same entity, they are extracted if they are on the same side from the *Action* entity as the one extracted or they are on the other side from the *Action* entity, but there is no additional *VP* verbal phrase in this part of the sentence.

It is very important if the recommendation consists of adverb phrases $B-ADVP$, after using the default chunking, one needs to split the sentence on that place or places, and use the proposed method separately on each part. Splitting the sentence on more parts is useful to catch more information that can be hidden if the sentence is not split.

In Algorithm 1 the pseudocode of the grammar and dictionary based named-entity linking is given, where the lines from 6 to 16 correspond to the first phase, while the lines from 17 to 20 correspond to the second phase of the method.

Algorithm 1 : Grammar and dictionary based named-entity linking.

1: Set the dietary recommendation, Φ
2: Obtain the *Chunks* vector by introducing the default chunking on Φ
3: Split Φ on the position of each adverb chunk into set Ψ
4: Set $counter = 1$
5: **for** each part in Ψ **do**
6: Set Φ=$\Psi[counter]$
7: Obtain the *Words* vector by using the word-level tokenization on Φ
8: Obtain the *Chunks* vector by introducing the default chunking on Φ
9: Obtain the X_{Chunks} matrix using the Eq. 1
10: Obtain the X_{Entities} matrix using the Eq. 2
11: Obtain the $X_{\text{Candidates}}$ matrix using the Eq. 3
12: Perform the first additional chunking
13: Perform the second additional chunking
14: Perform the third additional chunking
15: Recalucalte the X_{Chunks} matrix using the Eq. 1
16: Recalculate the $X_{\text{Candidates}}$ matrix using the Eq. 3
17: Select the *Action* entity by searching the predicate in the VP subtrees
18: Extract all entities of interest by using one of the two scenarios of the candidate selection phase
19: Extract the *Group* entity
20: Add the labels for (Subject, Predicate, Object) using *S*, *P*, *O* in order to catch the relations between the extracted entities
21: *counter* = *counter*+1
22: **end for**

5 EVALUATION AND RESULTS

We are interested in named-entity linking in the domain of dietary recommendations, so the entities (labels) of interest are the *Food* entity, the *Component* entity and the $Quantity/Unit$ entity. In addition, the *Action* entity and the *Group* entity are proposed by the method.

At the beginning, the dictionaries that will be used for each entity need to be defined.

For the $Quantity/Unit$ entity, an ontology, called Units of Measurements Ontology (UO) (Gkoutos

et al., 2012), is used. From it, the units together with their symbols are extracted. In addition, a list of units of measurements that are used for recipes, such as tablespoon, teaspoon, etc. are added.

For the *Component* entity, a combination of becas API (Nunes et al., 2013) and becas[chemicals] API (Campos et al., 2013) is used. becas[chemicals] is a web application and API for recognition and annotation of chemical compounds and drugs. It is a special branch of becas API focused on the identification of a large array of chemical substances. It uses machine-learning techniques, with an optimized feature set including orthographic, morphological, natural language processing, domain knowledge, and local context features.

For the *Food* entity a semantic tagger is used, known as USAS online English semantic tagger (Wilson and Thomas, 1997; McEnery and Wilson, 2001; Rayson et al., 2004), and the focus is on two categories. The first one is the category for terms related to Food and Farming, F. From it, three subcategories are used. The first one is the subcategory for terms related to food and food preparation, F_1, the second is for terms related to drinks and drinking, F_2, and the third subcategory is for terms related to cigarettes and drugs, F_3. The second category is for terms related to Life and Living things, L. From it two subcategories are used. The first one is for terms related to living creatures (e.g. non-human), L_2, and the second is for terms related to plants and plant-life, L_3.

5.1 Examples

In this section two examples are presented that illustrate how the proposed method works.

Let Φ_1 be the dietary recommendation "*People of any age who are Afro Americans should further reduce sodium intake to 300 mg per day.*".

In Table 2 the first phase of the grammar and dictionary based named-entity linking method for Φ_1 is presented. The *Tokens* column corresponds to the result of the word-level tokenization. The *POS tags* column corresponds to the result of the POS tagging. The *Chunk tokens* column corresponds to the result of the default chunking, where each chunk token is presented in *IOB* format and the beginning of each new chunk is marked with *. The *Food*, *Component* and *Quantity*/*Unit* columns correspond to the linking of each word to the knowledge reference bases that in our case are the USAS English semantic tagger, becas API and becas[chemicals] API, and the combination of the UO ontology and special recipe's metrics, respectively. In the column $Chunk_1$ *tokens* the result of the first additional chunking is presented, where only the new chunks formed by this chunking are presented in bold font. The new chunks formed by the second additional chunking are presented in bold font in the $Chunk_2$ *tokens* column, and in the $Chunk_3$ *tokens* column the result of the third additional chunking is presented and in this example nothing is changed by applying this chunking.

Then by using the Equation 3 the $X_{\text{Candidates}}$ matrix is calculated, where the rows correspond to the different chunks and the columns correspond to the entities, in our case *Food*, *Componenet* and *Quantity*/*Unit*. The $X_{\text{Candidates}}$ matrix has 6 rows (different chunks) and 3 columns. The *Food* column gives the set of the candidates for the *Food* entity, which in our example is an empty set because the dietary recommendation does not consist of food entities. The *Component* column gives the set of candidates for the *Component* entity, which in our case is a set with one element that is "sodium intake" identified by the row or the chunk that has nonzero element in the *Component* column. The *Quantity*/*Unit* column gives the set of the candidates for the *Quantity*/*Unit* entity, which in our example is a set of one element that is "300 mg per day".

After the first phase of the method, the recommendation Φ_1 is represented as undirected graph, where each chunk is connected only with its neighbours. In Figure 2 the graph representation of the recommendation Φ_1 is presented. Then the first step of the second phase is to select the initial node of the graph or the *Action* entity from where the search for all entities will start. To select the *Action* entity the parse

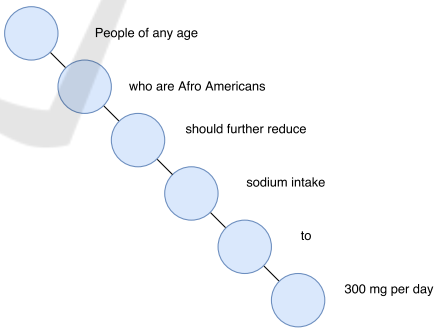

Figure 2: Graph representation of the recommendation Φ_1.

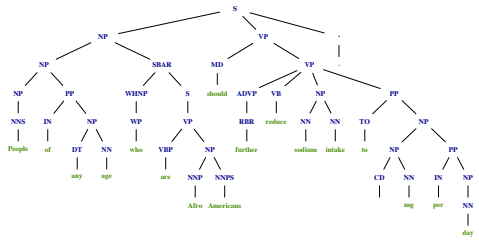

Figure 3: Parse tree of Φ_1.

Table 2: The first phase of the grammar and dictionary based named-entity linking method for Φ_1.

Tokens	POS tags	Chunk tokens	Food	Component	Quantity/Unit	$Chunk_1$ tokens	$Chunk_2$ tokens	$Chunk_3$ tokens
People	NNS	B-NP *	0	0	0	**B-NP** *	B-NP *	B-NP *
of	IN	B-PP *	0	0	0	**I-NP**	I-NP	I-NP
any	DT	B-NP *	0	0	0	**I-NP**	I-NP	I-NP
age	NN	I-NP	0	0	0	**I-NP**	I-NP	I-NP
who	WP	B-NP *	0	0	0	B-NP *	**B-NP** *	B-NP *
are	VBP	B-VP *	0	0	0	B-VP *	**I-NP**	I-NP
Afro	JJ	B-NP *	0	0	0	B-NP *	**I-NP**	I-NP
Americans	NNP	I-NP	0	0	0	I-NP	**I-NP**	I-NP
should	MD	B-VP *	0	0	0	B-VP *	B-VP *	B-VP *
further	RBR	I-VP	0	0	0	I-VP	I-VP	I-VP
reduce	VB	I-VP	0	0	0	I-VP	I-VP	I-VP
sodium	NN	B-NP *	0	**1**	0	B-NP *	B-NP *	B-NP *
intake	NN	I-NP	0	0	0	I-NP	I-NP	I-NP
to	TO	B-PP *	0	0	0	B-PP *	B-PP *	B-PP *
300	CD	B-NP *	0	0	0	**B-NP** *	B-NP*	B-NP *
mg	NN	I-NP	0	0	**1**	**I-NP**	I-NP	I-NP
per	IN	B-PP *	0	0	0	**I-NP**	I-NP	I-NP
day	NN	B-NP *	0	0	0	**I-NP**	I-NP	I-NP

tree of the recommendation Φ_1 is used. In Figure 3 we present the parse tree of the recommendation Φ_1. From it, the result of searching for predicate in the verbal phrases is the verb "should" from the *MD* subtree because it is closest to the root of the sentence. Further, the chunk that consists of the verb returned, "should further reduce", is selected as *Action* entity.

The last step of the second phase is to select all other important entities. By using the second scenario (since the *Action* entity is returned), we found one *Component* entity "sodium intake", one $Quantity/Unit$ entity "300 mg per day" and for the *Group* entity we obtained "People of any age", and "who are Afro Americans", while we did not find *Food* entity that is logical, because there are no food related terms in the recommendation. At the end, the labels for the $Subject$, $Predicate$ and $Object$ are added, ("People of any age", S_1), ("who are Afro Americans", S_1), ("should further reduce", P_1), ("sodium intake", O_1), and ("300 mg per day", O_1). The index of the labels indicates from which part of the sentence the entity is extracted. In this example is 1, because the recommendation does not contains any adverb chunks, so it is not split at the beginning.

In the second example let Φ_2 be the following dietary recommendation "*The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women, while for men and women over 50 it is 30 g and 21 g per day, respectively, due to decreased food consumption.*". The difference with the previous example is that this recommendation consists of two adverb chunks, so it needs to be split. If the recommendation is not split, than the extracted entities are "The recommended intake for adults", and "50 years" as *Group* entities, "is set" and "is" as *Action* entities, "decreased food consumption" as *Food* entity, and "38 g for men", and "25 g for women" as $Quantity/Unit$ entities, and the information for men and women over 50 is still hidden and it is not extracted.

In such case, the recommendation is split on the position of each adverb chunk. In this recommendation, there are two adverb chunks, "while" and "respectively", so we split it in three parts, "*The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women.*", "*For men and women over 50 it is 30 g and 21 g per day.*", and "*Due to decreased food consumption.*".

Then the proposed method is used on each part of the recommendation obtained after splitting. For the first part, the extracted entities are ("The recommended intake for total fiber for adults", S_1), ("50 years", S_1), and ("younger", S_1) as *Group* entities, ("is set", P_1) as *Action* entity, and ("38 g for men", O_1), and ("25 g for women", O_1) as $Quantity/Unit$ entities. By applying the method on the second part of the recommendation, the extracted terms are ("For men and women over 50", S_2), and ("it", S_2) as *Group* entities, ("is", P_2) as *Action* entity, and ("30 g", O_2), and ("21 g per day", O_2) as $Quantity/Unit$ entities. For the third part of the recommendation, only one extracted term exists ("decreased food consumption",

Table 3: Knowledge extraction of 15 dietary recommendations.

	Recommendation	Group	Action	Food	Component	Quantity/Unit
1.	Good sources of magnesium are: fruits or vegetables, nuts, peas and beans, soy products, whole grains and milk.	-	are (P_1)	fruits or vegetables, nuts, peas and beans (O_1) soy products (O_1) whole grains and milk (O_1)	Good sources of magnesium (S_1)	-
2.	The RDAs for Mg are 300 mg for young women and 350 mg for young men.	-	are (P_1)	-	The RDAs for Mg (S_1)	300 mg for young women (O_1) 350 mg for young men (O_1)
3.	Increase potassium by ordering a salad, extra steamed or roasted vegetables, bean-based dishes fruit cups, and low-fat milk instead of soda.	-	-	salad (S_1) extra steamed or roasted vegetables (S_1) fruit cups (S_1) law-fat milk (S_1)	Increase potassium (S_1)	-
4.	Babies need protein about 10 g a day.	Babies (S_1)	need (P_1)	-	protein (O_1)	10 g a day (O_1)
5.	1 teaspoon of table salt contains 2300 mg of sodium.	-	contains (P_1)	-	table salt (S_1) sodium (O_1)	1 teaspoon (S_1) 2300 mg (O_1)
6.	Milk, cheese, yogurt and other dairy products are good sources of calcium and protein, plus many other vitamins and minerals.	-	are (P_1)	Milk, cheese, yogurt and other diary products (S_1)	good sources of calcium and protein (O_1) many other vitamins and minerals (O_1)	-
7.	Breast milk provides sufficient zinc, 2 mg/day for the first 4-6 months of life.	-	provides (P_1)	Breast milk (S_1)	sufficient zinc (O_1)	2 mg/day for the first 4-6 months of life (O_1)
8.	If you're trying to get more omega-3, you might choose salmon, tuna, or eggs enriched with omega-3.	you (S_1)	might choose (P_1)	salmon, tuna, or eggs (O_1)	more omega-3 (S_1)	-
9.	If you need to get more fiber, look to beans, vegetables, nuts and legumes.	more fiber (S_1)	look (P_1)	beans, vegetables, nuts, and legumes (O_1)	-	-
10.	Eating foods high in vitamin C and iron can reduce the absorption of ingested nickel.	-	can reduce (P_1)	Eating foods (S_1)	vitamin C and iron(S_1) the absorption of ingested nickel (O_1)	-
11.	The body of a 76 kg man contains about 12 kg of protein.	-	contains (P_1)	-	protein (O_1)	The body of a 76 kg man (S_1) about 12 kg (O_1)
12.	Excellent sources of alpha-linolenic acid, ALA, include flaxseeds and walnuts.	-	include (P_1)	flaxseeds and walnuts (O_1)	Excellent sources of alpha-linolenic acid (S_1) ALA(S_1)	-
13.	The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women, while for men and women over 50 it is 30 g and 21 g per day, respectively, due to decreased food consumption.	The recommended dietary intake for total finer for adults (S_1) 50 years (S_1) younger (S_1) for men and women over 50 (S_2) it (S_2)	is set (P_1) is (P_2)	decreased food consumption(S_3)	-	38 g for men (O_1) 25 g for women (O_1) 30 g (O_2) 21 g per day (O_2)
14.	I'm good at tennis.	-	-	-	-	-
15.	Your hat looks very nice.	-	-	-	-	-

S_3).

In this example the term "fiber" is not extracted as *Component* entity, but this happens because "fiber" is not annotated as chemical term using the becas API. If it would be added into the dictionary for the *Component* entity, then it will be annotated and extracted as *Component* entity.

5.2 Experiment

In order to evaluate the method, a collection of 50 dietary recommendations provided by the World Health Organization and the U.S. National Library of Medicine is analyzed. The collection of 50 dietary recommendations can be requested from the authors for future analysis. In addition, 10 sentences that are not related with dietary recommendations are added.

By using the proposed method, for 47 out of 50 recommendations all the useful knowledge is extracted. For 2 out of 50 recommendations only the nutrient component is not extracted, which is "fiber" because this term is not annotated as chemical term using the becas API. However we have this information in the *Group* entity. Also, for 1 of them the "sodas" is not extracted as *Food* entity because it is not annotated using the USAS semantic tagger, but the term is in the *Group* entity. Having these terms in the *Group* entity is still a benefit because the information about them is extracted and can be later modified by the human experts.

By using the method for the 10 sentences that are not dietary recommendations, only the *Group* entity and the *Action* entity can be extracted because the $X_{\text{Candidates}}$ matrix contains only zero elements, so we do not continue with the extraction.

In Table 3 we present the results obtained for 15 randomly selected dietary recommendations.

6 CONCLUSION

In this paper a grammar and dictionary based named-entity linking method is presented that can be used for knowledge extraction of evidence-based dietary recommendations. The method works with dietary recommendation presented in one sentence. It consists of two phases. The first one is a mix between the entity detection and determination of a set of candidates for each entity, while the second phase is a candidate selection. The focus is on food related entities, nutrient related entities, and quantity/unit entities. The method is evaluated using dietary recommendations provided by the World Health Organization and the U.S. National Library of Medicine.

To the best of our knowledge, this is the first named-entity linking method that is focused on entities related with dietary recommendations. In the absence of labeled data needed for the current state of the art machine learning approaches, the benefit of this method is that can easily extracted the entities from unstructured data. In the future we are planning to extend this method to work with text paragraphs and to extract all possible entities of interest.

In addition we provide a named-entity linking method that can be used also in other domains, by using appropriate dictionaries for the entities in those domains.

REFERENCES

A.Voutilainen (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232.

Blanco, R., Boldi, P., and Marino, A. (2015). Using graph distances for named-entity linking. *Science of Computer Programming*.

Campos, D., Matos, S., and Oliveira, J. L. (2013). Chemical name recognition with harmonized feature-rich conditional random fields. In *BioCreative Challenge Evaluation Workshop*, volume 2, page 82.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

EFSA ((accessed February 18, 2016)). *European Food safety Authority*. https://www.efsa.europa.eu/.

Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. (2012). The units ontology: a tool for integrating units of measurement in science. *Database*, 2012:bas033.

Hachey, B., Radford, W., Nothman, J., Honnibal, M., and Curran, J. R. (2013). Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150.

Han, X., Sun, L., and Zhao, J. (2011). Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 765–774. ACM.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

McEnery, T. and Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh University Press.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

Nelson, R. J. (1955). Karnaugh m.. the map method for synthesis of combinational logic circuits. transactions of the american institute of electrical engineers, vol. 72 part i (1953), pp. 593–598. *The Journal of Symbolic Logic*, 20(02):197–197.

Nunes, T., Campos, D., Matos, S., and Oliveira, J. L. (2013). Becas: biomedical concept recognition services and visualization. *Bioinformatics*, page btt317.

Rayson, P., Archer, D., Piao, S., and McEnery, A. (2004). The ucrel semantic analysis system.

Rusu, D., Dali, L., Fortuna, B., Grobelnik, M., and Mladenic, D. (2007). Triplet extraction from sentences. In *Proceedings of the 10th International Multiconference" Information Society-IS*, pages 8–12.

Santorini, B. (1990). Part-of-speech tagging guidelines for the penn treebank project (3rd revision).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

Taylor, A., Marcus, M., and Santorini, B. (2003). The penn treebank: an overview. In *Treebanks*, pages 5–22. Springer.

Tian, Y. and Lo, D. (2015). A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports. In *Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on*, pages 570–574. IEEE.

Vorster, H., Love, P., and Browne, C. (2001). Development of food-based dietary guidelines for south africa: the process. *S Afr J Clin Nutr*, 14(3).

Wilson, A. and Thomas, J. (1997). Semantic annotation. *Corpus Annotation. Longman, London*.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS