A SEMI-AUTOMATIC SYSTEM FOR CLASSIFYING AND DESCRIBING FOODS ACCORDING TO FOODEX2

Tome Eftimov [1),2)], *Gordana Ispirova* [1),2)], *Barbara Koroušić Seljak* [1)], *Peter Korošec* [1),3)]
[1)] Computer Systems Department (Jožef Stefan Institute), Ljubljana, Slovenia,
tome.eftimov@ijs.si, gordana.ispirova@ijs.si, barbara.korousic@ijs.si, peter.korosec@ijs.si
[2)] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3)] Faculty of Mathematics, Natural Science and Information Technologies, Koper, Slovenia

Abstract–In this paper, we present results of the evaluation of a semi-automatic system for classifying and describing foods according to FoodEx2 using datasets from three European countries. The proposed system is an integration of methods from machine learning, natural language processing and probability theory. It obtained an accuracy of 91% for the classification part and 78% for the description part. The usage of the system can be a link between food consumption and food composition data as the transformation from food intake into nutrient intake can be automatically made.

Keywords: *food standardization, food matching, FoodEx2, food classification, food description*

1. INTRODUCTION

Nowadays, dietary information may be mentioned in text in multiple ways, using phrases with a variety of structures. The alternative ways in which the same concept may be mentioned in food information systems creates a problem of the automatic integration of information [1]. For example, one problem appears when we are interested in linking the food intake information to food composition data. The problem is that the name of the same food can be different regarding to the language or to different ways of how people express themselves. For this reason, European Food Safety Authority (EFSA), has provided a standardized food classification and description system called FoodEx2. It uses facets to describe food properties and aspects from various perspectives. The usage of the system makes easier to compare food consumption data from different sources and perform more detailed types of data analysis. However, a lot of food composition data and food consumption data, which need to be linked, are lacking in FoodEx2 classifiers and descriptors. This happens because the process of classification and description needs to be donemanually, which is a

time-consuming task that also requires a good knowledge of the system and a good knowledge on food.

FoodEx2 consists of three types of food categories representing three different levels in the food chain, which involve an increasing level of food processing i.e., moving from raw commodities to derivatives to composite foods. The food categories are defined in FoodEx2 technical report [2].

FoodEx2 includes two types of terms: list terms and facet descriptors. List terms are represented by a code, for example, the A0C75 is the code for "salmon". Facet descriptors are elements of additional information included in or added to the list terms, each providing different options to describe a particular aspect of a food category, such as treatments received, production method, fat content, and qualitative information.

In this study, we used a recently proposed semi-automatic system, called StandFood, to standardize foods according to FoodEx2 [3]. The system consists of three parts. The first identifies what type of food is being analyzed. This is the classification part that involves a machine learning (ML) approach [4]. The second describes the food using natural language processing (NLP) [5] combined with probability theory, which results with the list term or FoodEx2 code for the food. The third combines the result from the first and the second part by defining post-processing rules in order to improve the result for the classification part.

2. EXPERIMENTAL

For the classification part of the system, 5,416 FoodEx2 names were used as the training instances. They are pre-processed by removing numbers and punctuations and stemming.

Then, the document-term matrix is built, so it can be used for feature selection. Additional features to the problem are added to the selected features, which are the number of nouns found in the food description name, number of adjectives, number of verbs, and the length of food description name. Different classifiers and ensemble learning are used to obtain a model that can be further used to predict the food category of new unseen food items [3]. The selected model is an ensemble of four algorithms: Support Vector Machine (SVM), Random forest (RF), Maximum entropy (Maxent), and Boosting, combined using a majority vote strategy.

After identifying the category of the food item, it is then necessary to describe it using the FoodEx2 code, the list term. For the description part, for each food item that needs to be described according to FoodEx2, its English name is used. The name is pre-processed by converting it to lowercase letters. Part-Of-Speech (POS) tagging is used to extract its nouns, adjectives, and verbs [6]. The extracted sets are further transformed using lemmatization [7]. Using the extracted nouns, the FoodEx2 data is searched for the names that consist of at least one of the extracted nouns. Then, the resulting list (subset) is pre-processed by converting each food item name to lowercase letters, applying POS tagging to extract the nouns, adjectives, and verbs, and using lemmatization for the extracted sets. Then, the food item that needs to be described according to FoodEx2 is matched with each food item in the resulting list and a weight which is the probability obtained to find a similarity between the extracted sets of nouns, adjectives and verbs, is assigned on each matching pair. Finally, the pair with the highest weight is the most relevant one, so it is returned together with its food category from FoodEx2.

Results of the description part are used to improve the performance of the classification part. Because for each food item the most relevant item from FoodEx2 is returned, the category of the returned food item can be used for post-processing rules. To do this, StandFood applies the following four rules:

In the first rule, processes that change the nature of a food product are used. Their definitions are in the technical report of FoodEx2 and include, for example, *canning*, *smoking*, *frying*, and *baking*. Then, for each process in the list of processes, lemmatization is applied to avoid different word forms of food items names. So, if a food item is classified as raw (r) using the StandFood classification model but its set of adjectives and verbs (their lemmas) consists of at least one cooking processthat change its nature, it is automatically changed to a derivative (d).

In the second rule, if a food item is classified, as being either raw (r) or a derivative (d) and the result from the description part of StandFood is that the most relevant food item is a composite food (s) or (c), it is automatically changed to a composite food.

In the third rule, if a food item is classified as a simple composite food (s) and the result from the description part of the StandFood system is that the most relevant food item is an aggregated composite food (c), it is automatically changed to an aggregated composite food (c).

The fourth rule works in reverse, by changing an aggregated composite food (c) to a simple composite food (s).

The first evaluation of the proposed system was made using food items from one European country. In this study, an additional evaluation of the system was made using datasets from three European countries, or 682 food items, where foods were already classified and described using FoodEx2 codes. In the dataset, each food item is represented by a food name and a FoodEx2 code, which is manually added by a human expert. For one country, the manually coding was made by searching FoodEx2 hierarchy; while for the other two countries the manual coding was made using the search browser provided by FoodEx2. StandFood was then used, first to provide the food category to which the item belongs, and second to describe it using FoodEx2 code. This was then compared with the food category and the code that was added manually.

3. RESULTS AND DISCUSSION

StandFood was used to find the food category and the list term for 682 food items. The results obtained when the manual coding was performed by looking up the relevant categories and list terms in the hierarchical list are comparable with the results obtained when the manually coding was performed using the search option provided in the FoodEx2 system. For the classification part, the overall result is

91%, while for the description part the overall result is 78%.

To become familiar with the results, Table 1 gives the results from the StandFood classification part of eight randomly selected but correctly classified instances, two per food category.

Table 1. Correctly classified instances by the StandFood classification part

Food item	ategory
Kale, raw	r
Salmon, fresh	r
Flour, durum wheat	d
Cheese, soft, mascarpone	d
Wholemeal rye bread	s
Shandy (beer +lemonaide)	s
Pastry cream puff	c
Soup, fish, canned	c

After classifying food items, the next step is to describe them using FoodEx2 list term. Table 2 gives the list terms for eight randomly selected instances. Food item represents the name of the food item in the data set. List term (StandFood) is the FoodEx2 code of the most relevant match found by the StandFood. List term (manual) is the FoodEx2 code that was manually assigned to that food item by a human expert. Table 3 gives the relevant FoodEx2 food items obtained by StandFood for the same eight instances.

Table 2. List terms for eight randomly selected food items

Food item	List term (StandFood)	List term (manual)
Banana dried	A01MJ	A01MJ
Omlette with mushrooms	A03YR	A03YR
Brown sauce (gravy, lyonnais sauce)	A043Z	A043Z
Sugar, white	A032J	A032H
Juice, apricot	A039N	A03BC
Hard cheese: kefalotyri	A02YT	A02YE
Coffee with milk or cappuccino, unsweetened	A03KG A03KH	A03KG
Sweets (candy), average	A034X A035L	A034X

Table 3. Relevant FoodEx2 food items for eight randomly selected instances obtained from the StandFood description part

Food item	FoodEx2 food item (StandFood)
Banana dried	Dried bananas
Omlette with mushrooms	Ommlete with msuhrooms
Brown sauce (gravy, lyonnais sauce)	Continental European brown cooked sauce gravy
Sugar, white	White sugar
Juice, apricot	Juice apricot
Hard cheese: kefalotyri	Cheese kefalotyri
Coffee with milk or cappuccino, unsweetened	Coffee with milk or cream; Coffee drink cappuccino
Sweets (candy), average	Hard candies; Jelly candies

Using tables 2 and 3, for the first three food items, the FoodEx2 list term obtained by StandFood is the same as the Foodex2 list term that is manually assigned, even in the case when there is variability in the food item name. For the next three food items, the FoodEx2 list term obtained by StandFood and manually assigned list term differ. For two of them, for “hard cheese: kefalotyri” and “sugar, white”, the list terms are very close in the FoodEx2 hierarchy. The manually assigned list terms for both food items, A032H and A02YE, are parents (more broad terms) for the StandFood obtained list terms, A039N and A02YT, respectively. For “juice, apricot”, the StandFood obtained list term is for “juice apricot” that has already existed in FoodEx2, while the manually assigned list term is for “nectar apricot”. For the last two food items, the manually assigned Foodex2 list term is one of the list terms returned by StandFood.

In this case,there is several food items from the FoodEx2 data that have the same weight, which is the highest weight obtained, so they are all returned as relevant matches. The user should select which of the returned matches is the most relevant one.

After obtaining the result from the StandFood description part, the food category of the most relevant match is further combined with the obtained food category from the StandFood classification part. They are combined in post-processing rules in order to improve the performance of the classification part.

As an example, post-processing of four randomly selected foods was carried out in Table 4.

Table 4. Food categories for four food items after post-processing rules

Food item	Food category (classificaiton)	ood category (post-processing)
Coffee beans roasted	r	d
Potato boiled	r	d
Croissant	r	c
Croissant with jam	s	c

The main benefit of using StandFood is that it can be a link between food consumption and food composition data. If both datasets consist of FoodEx2 code the transformation from food intake into nutrient intake can be automatically made. Using the trained classification model on the FoodEx2 data, there are some weaknesses when it is applied on new instances. Despite the weaknesses, it can still be used because the StandFood classification part includes post-processing rules in order to improve its performance. Also, it is seen that there are some cases when the StandFood gives more a relevant FoodEx2 code than a human expert. This happens because human experts manually code food items with facet descriptors or use more broad terms from FoodEx2 hierarchy.

According to the time needed to code the new 682 food items, in the case of StandFood minutes (five minutes) are required because the current version is programmed in a sequential way. Parallel programming would allow the same task to be carried out in seconds. A human expert takes on average ten minutes per item, or 6820 min for 682 food items, or 14 days if a humane expert works eight hours per day.

However, the result from standfood needs to be checked and in the case of several options to choose the correct one. For 682 food items, through checking and choosing the correct result when several options were given took one working day**.**

4. CONCLUSIONS

This study presents a semi-automatic system for classifying and describing food items according to the Foodex2 standard. In practice, the system can be used to find missing FoodEx2 codes for food composition data and food consumption data that are basic resources that need to be linked and combined for dietary assessment methods.

ACKNOWLEDGMENTS

This work is supported by the project ISO-FOOD, which received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 621329 (2014-2019). This work is also part of the National Dietary Survey in compliance with the EFSA EU Menu methodology (SI.Menu 2017-18), supported by EFSA (Contracts OC/EFSA/DATA/2014/02-LOT1-CT02,OC/EFSA/DATA /2014/02-LOT2-CT03). This work is also supported by the project RICHFIELDS, which received funding from the European Union's Horizon 2020 research and innovation programme under grant number 654280. The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0098).

REFERENCES

[1] Gurinovic, M.; Mileševic, J.; Kadvan, A.; Djekic-Ivankovic, M.; Debeljak-Martacic, J.; Takic, M.; Nikolic, M.; Rankovic, S.; Finglas, P.; Glibetic, M. Establishment and advances in the online Serbian food and recipe data base harmonized with EuroFIRTM standards. Food Chem. 2016, 193, 30–38.

[2] European Food safety Authority. The Food Classification and Description System FoodEx2, 2nd ed.; European Food safety Authority: Parma, Italy; Available
online: https://www.efsa.europa.eu/ (accessed on 17 February 2017).

[3] Eftimov, T.; Korošec, P.; Koroušić Seljak, B. "StandFood: Standardization of Foods Using a Semi-Automatic System for Classifying and Describing Foods According to FoodEx2." Nutrients 9.6.2017: 542.

[4] Michalski, R.S.; Carbonell, J.G.; Mitchell, T.M. Machine Learning: An Artificial Intelligence Approach; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.

[5] Chowdhury, G.G. Natural language processing. Annu. Rev. Inf. Sci. Technol. 2003, 37, 51–89.

[6] Voutilainen, A. Part-of-speech tagging. In The Oxford Handbook of Computational Linguistics; Oxford University Press Inc.: New York, NY, USA, 2003; pp. 219–232.

[7] Plisson, J.; Lavrac, N.; Mladenic, D. A rule based approach to word lemmatization. Proc. IS 2004, 83–86.