On the lexical coverage of some resources on Italian cooking recipes

Alessandro Mazzei
Università degli Studi di Torino
Corso Svizzera 185, 10149 Torino
mazzei@di.unito.it

Abstract

English. We describe an experiment designed to measure the lexical coverage of some resources over the Italian cooking recipes genre. First, we have built a small cooking recipe dataset; second, we have done a qualitative morpho-syntactic analysis of the dataset and third we have done a quantitative analysis of the lexical coverage of the dataset.

Italian. *Descriviamo un esperimento per valutare la copertura lessicale di alcune risorse sul genere delle ricette da cucina. Primo, abbiamo costruito un piccolo dataset di ricette. Secondo, ne abbiamo eseguito un'analisi qualitativa di sulla morfo-sintassi. Terzo, ne abbiamo eseguito un'analisi quantitativa della copertura lessicale.*

Introduction

The study reported in this paper is part of an applicative project in the field of nutrition. We are designing a software service for Diet Management (Fig. 1) that by using a smartphone allows one to retrieve, analyze and store the nutrition information about the courses. In our hypothetical scenario the interaction between the man and the food is mediated by an intelligent recommendation system that on the basis of various factors encourages or discourages the user to eat that specific course. The main factors that the system needs to account for are: (1) the diet that you intend to follow, (2) the food that have been eaten in the last days and, (3) the nutritional values of the ingredients of the course and its specific recipe. Crucially, in order to extract the complete salient nutrition information from a recipe, we need to analyze the sentences of the recipe. To allow the execution of this information extraction task we intend to use a syntactic parser together with a semantic interpreter. However, we intend to use both a *deep* interpreter, which have showed to have good performances in restricted domains (Fundel et al., 2007; Lesmo et al., 2013), as well as a *shallow* interpreter, that are most widely used in practical applications (Manning et al., 2008).

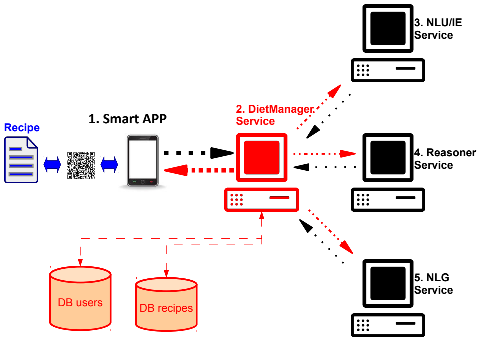

Figure 1: The architecture of the diet management system.

In order to optimize the information extraction task we need to evaluate the specific features of the applicative domain (Fisher, 2001; Jurafsky, 2014). In the next Sections we present a preliminary study towards the realization of our NLP system, i.e. a linguistic analysis of the Italian recipes domain.

1 Data set construction

The construction of a linguistic resource includes three main steps, i.e. collection, annotation and analysis of linguistic data. Since all these steps are very time-consuming, it is usual perform first off all tests on a preliminary small dataset. As a case study we selected three versions of the same recipe, that of the "caponata" (a Sicilian course consisting of cooked vegetables), respectively extracted from a WikiBook (210 tokens,

	Tok./Sent.	N/V	Con/Fun words
WikiBook	11.8	2.3	1.7
Cucchiaio	16.6	2.8	1.5
Cuochi	21.7	1.9	1.2

Table 1: The rates of the number of the tokens for sentences, of the number of nouns respect to verbs, of the number of the content words w.r.t function words.

15 sentences, *WikiBook* in the following) (Wikibooks, 2014), and from two famous Italian cooking books: "Il cucchiaio d'argento" (399 tokens, 23 sentences, *Cucchiaio* in the following) (D'Onofrio, 1997), and "Cuochi si diventa" (355 tokens, 16 sentences, *Cuochi* in the following) (Bay, 2003). The corpus obtained consists of 964 tokens corresponding to 54 sentences. The application of treebank development techniques to this very limited amount of data is only devoted to a preliminary qualitative evaluation of the feasibility of the information extraction task and the detection of the major difficulties that can be expected in the further development of our work. For what concern the collection of texts, the selection of data from three different books will have some impact on the further steps. In particular, this allows us to find different lexical choices in the three data sets, or different exploitation of specific linguistic constructions, such as passive versus active clauses, or different frequency of specific verbal forms, such as imperative versus present. Moreover, we can find different structures used to describe recipes or, in other words, different text styles within the cooking text genre.

In Table 1 we reported some statistics about the corpus. The number of tokens for sentence and the rate between content and function words reveal that *WikiBook* uses a simpler register with respect to the other sources. The style used by *Cuochi*, that is similar to a novel and does not follow the standard ingredients-methods template (Fisher, 2001), is revealed by the high number of tokens for sentence.[1]

2 Morpho-syntactic analysis

Following a typical strategy of semi-automatic annotation, i.e. automatic annotation followed by manual correction, for the annotation of our small dependency treebank we applied on the preliminary dataset two pipelines which integrate morphological and syntactic analysis, i.e. TULE (Turin University Linguistic Environment) (Lesmo, 2007) and DeSR (Dependency Shift Reduce) (Attardi, 2006). The exploitation of two different systems allows the comparison of the different outputs produced and the selection of the best one. Both TULE and DeSR have been tuned on a balanced corpus that does not contain recipes.

As far as the morphological analysis is concerned, first of all we have to observe that each error in the Part of Speech tagging (PoS, $1.7 - 3.2\%$ for TULE), such as the erroneous attribution of the grammatical category to a word (e.g. Verb rather than Noun), has an effect on the following analysis. For instance, it makes impossible to build a syntactic tree for some sentence or to recovery a meaning for some word in the semantic database. Because of this motivation, we started our error recovery process from the morphological annotation.

As far as the syntactic analysis is concerned, the performance of the parsers adopted are qualitatively similar even if the errors can vary. The problems more frequently detected are related to the sentence splitting which can be solved by using a pre-processing step. These problems are rare in *Cuochi* and mainly found in *Cucchiaio* or *WikiBook*, where the recipes are organized by a set of titles according to a sort of template (cf. (Fisher, 2001)), including e.g. the name of the recipe, "Ingredienti", "Ricetta", "Per saperne di più". This confirms that the selected books adopt a different style in the description of recipes also within the same text genre represented by cooking literature.

3 Lexical coverage experiment

With the aim to extract information from recipes (Maeta et al., 2014; Walter et al., 2011; Amélie Cordier, 2014; Shidochi et al., 2009; Haoran Xie and Lijuan Yu and Qing Li, 2010; Druck, 2013), a key issue regards the coverage of the lexicon. In order to extract the nutrition values from a specific recipe, we need to map the words contained into the recipe to a semantic organized repositories of lexical knowledge. A number of lexical resources are specialized on one specific domain while others resources are more general and, often, are automatically extracted from semi-structured

[1] For example, an newspaper section of the Turin University Treebank has $\sim$ 25 tokens for sentence (Bosco et al., 200).

resources (Hovy et al., 2013). In order to explore the automatic extraction of information from Italian recipes, we designed an experiment that uses both the types of resources.

In our experiment we have used 4 distinct Italian computational lexicons: 1 specialized lexicon, i.e. AGROVOC (FAO, 2014), and 3 general lexicons, i.e. MultiWordNet, BabelNet, Universal-Wordnet. AGROVOC is a specialized lexicon, that is a controlled multi-language vocabulary, developed in collaboration with the FAO, covering a number of domains related to food, as nutrition, agriculture, environment, etc. It contains $40,000$ concepts organized in a hierarchies, that express lexical relations among concepts, as "narrow terms". Each concept is denoted by a number, and can be linked by different lexical items (terms) in different languages. AGROVOC is formalized as a RDF linked dataset but it is also available for download in various formats.[2] A notable feature of AGROVOC is the direct connection with other knowledge repository: in particular it is connected with DBpedia (Bizer et al., 2009), that often contains explicit annotation of the nutrition values.

MultiWordNet, BabelNet and UniversalWordnet are three general computational lexicons related to WordNet, that is a large lexical database of English (Miller, 1995; Fellbaum, 2005). Nouns, verbs, adjectives and adverbs are organized into sets of synonyms (synsets), each one denoting a distinct concept. Synsets are interlinked by means of semantic and lexical relations as ISA relation or hyperonymy relation. MultiWordNet is a lexical database in which an Italian WordNet is strictly aligned with WordNet[3]. The Italian synsets ($\sim 30,000$, that are linked by $\sim 40,000$ lemmas) are created in correspondence with the WordNet synsets and the semantic relations are imported from the corresponding English synsets (Pianta et al., 2002). BabelNet is a multilingual lexicalized ontology automatically created by linking Wikipedia to WordNet (Navigli and Ponzetto, 2012). The integration is obtained by an automatic mapping and by using statistical machine translation. The result is an "encyclopedic dictionary" that provides concepts and named entities lexicalized in many languages, among them Italian. In this work we used BabelNet 1.1 consisting of 5 millions of concepts linked by 26 millions of word. UniversalWordNet is an automatically constructed multilingual lexical knowledge base based on WordNet (de Melo and Weikum, 2009). Combining different repositories of lexical knowledge (e.g. wikipedia), UniversalWordNet cosists of $1,500,000$ lemmas in over 200 languages. Note that the direct connections of UniversalWordNet and BabelNet towards wikipedia allows one to access to the nutrition values of foods since they are often represented in wikipedia.

In order to analyze and compare the possible use of these Italian lexical resources for information extraction, we performed a Named-Entity Recognition (NER) experiment. We introduced three semantic entities that are particularly relevant for the recipe analysis: FOOD, PREP (preparation), Q/D (quantity and devices). We mark with the FOOD label the words denoting food, e.g. melanzana (*aubergine*), pomodoro (*tomato*), sale (*salt*); we mark with PREP words denoting verbs that are involved with the preparation of a recipe, e.g. tagliare (*to cut*), miscelare (*to mix*), cuocere (*to cook*); we mark with Q/D words expressing quantities, e.g. minuti (*minutes*), grammi (*grams*) or denoting objects that are related with the recipe preparation, e.g. cucchiaio (*spoon*), coltello (*knife*). By using these three name entity categories, we annotated the three caponata recipes. In the columns "Tok." (tokens) of the Tables 2-3-4 are reported the number of words for each category.

We performed two distinct experiments for lexical coverage. The first experiment concerns AGROVOC, the second experiment concerns MultiWordNet, BabelNet and Universal-WordNet. In the first experiment we count the number of entities that can be retrieved by a straight search in AGROVOC for each name entity category: we search for the word form and for the corresponding lemma too. The columns AgrVoc-TP (true positives) of the Tables 2-3-4, report the number of retrieved tokens for each category, and the columns AgrVoc-rec report the corresponding coverage. In this experiment there are no "false positives", i.e. all the elements retrieved belongs to a meaningful categories (in other word precision is 100%). A first consideration regards the very low scores obtained on the PREP and Q/D categories. This fact could suggest that AGROVOC lexicon is not enough gen-

[2] In the experiment we used the SQL version of AGROVOC.

[3] MultiWordNet is natively aligned with WordNet 1.6. However we adopt a pivot table in order to use WordNet 3.0.

eral to be used for recipe analysis. A deeper analysis explains also the low score obtained on the FOOD category. Many of the terms are present in AGROVOC only in the plural form: for instance AGROVOC contains the entry "pomodori" (*tomatoes*) but does not contain "pomodoro" (*tomato*). Moreover, many food do not have a generic lexical entry: for instance AGROVOC contains the entry "peperoni verdi" (*green peppers*) but does not contain "peperoni" (*peppers*). However, the best scores for this experiment has been obtained on *WikiBook*, that is on the simplest recipe.

The second experiment, that involves MultiWordNet, BabelNet and UniversalWordNet, is more complex. We use a naive *Super-Sense Tagging algorithm* (NaiveSST) for the NER task. SST consists of annotating each significant entity in a text (nouns, verbs, adjectives and adverbs) within a general semantic tag belonging to the taxonomy defined by the WordNet lexicographer classes, that are called *super-senses* (Ciaramita, 2003). The lexicographer classes are 44 general semantic categories as "location", "food", "artifact", "plant", etc. The NaiveSST algorithm is very simple:

```
foreach content word in the sentence do
    Retrieve all the synsets corresponding to
    the word from MultiWordNet, BabelNet,
    UniversalWordNet
    foreach super-sense of a synset do
        if the super-sense is food or plant or
        animal then
            assign the label FOOD to the word
        end
        if the super-sense is quantity or artifact
        then
            assign the label Q/D to the word
        end
        if  the super-sense is creation or
        change or contact then
            assign the label PREP to the word
        end
    end
end
```

Algorithm 1: The NaiveSST algorithm.

	Tok.	AgrVoc		NaiveSST			
		TP	rec%	TP	FP	pre%	rec%
FOOD	37	23	62.2	35	5	87.5	94.6
PREP	19	1	5.3	15	8	65.2	79.0
Q/D	15	6	40.0	8	10	44.4	53.3
TOT.	71	30	42.3	58	23	71.6	81.7

Table 2: The results of the lexical semantic coverage experiment on the "WikiBook" version of the caponata recipe.

	Tok.	AgrVoc		NaiveSST			
		TP	rec%	TP	FP	pre%	rec%
FOOD	61	35	57.4	55	10	84.62	90.2
PREP	49	4	8.2	35	12	74.5	71.4
Q/D	31	1	3.2	27	10	73.0	87.1
TOT.	141	40	28.4	117	42	73.6	83.0

Table 3: The results of the lexical semantic coverage experiment on the "Cucchiaio d'argento" version of the caponata recipe.

The columns NaiveSST-TP (true positives), NaiveSST-FP (false positives) of the Tables 2-3-4 report the number of correct/uncorrect labels for each category, while the NaiveSST-pre and NaiveSST-rec columns report the corresponding precision and recall. In contrast with the previous experiment, the best scores here has been obtained on *Cuochi*. Indeed, the novel-style of *Cuochi* gives better results on the PoS tagging ($\sim 1.7\%$) and, as a consequence, on the correct lemmatization of the words. Also in this experiment the most difficult category is Q/D: this low value is related to the lemmatization process too. Often the lemmatizer is not able to recognize the correct lemma, e.g. "pentolino" (*small pot*) or "″" (*seconds*).

	Tok.	AgrVoc		NaiveSST			
		TP	rec%	TP	FP	pre%	rec%
FOOD	45	27	60.0	43	11	79.6	95.6
PREP	52	2	3.9	49	4	92.5	94.2
Q/D	43	3	7.0	32	26	55.2	74.4
TOT.	140	32	22.9	124	41	75.15	88.6

Table 4: The results of the lexical semantic coverage experiment on the "Cuochi si diventa" version of the caponata recipe.

Conclusions

In this paper we presented a preliminary study on cooking recipes in Italian. The qualitative analysis emphatizes the importance of the sentence splitter and of the PoS tagger for a correct morpho-syntactic analysis. From the quantitative lexical coverage analysis we can draw a number of speculations. First, there is a great linguistic variation among cookbooks. Second, general lexical resources outperform domain specific resources with respect to lexical coverage. Third, the lemmatization can improve the recall of the algorithm with respect to the lexical resource.

Acknowledgments

This work has been partially supported by the project MADiMAN, partially funded by Regione Piemonte, Innovation Hub for ICT, POR FESR 2007/2013 - Asse I - Attività I.1.3.
`http://di.unito.it/madiman`

We thank Cristina Bosco and Manuela Sanguinetti for their precious help in the linguistic analysis of the recipes.

References

Amélie Cordier. 2014. 4th Computer Cooking Contest, An event of ICCBR 2011. http://liris.cnrs.fr/ccc/ccc2011/doku.php.

G. Attardi. 2006. Experiments with a multilanguage non–projective dependency parser. In *Proceedings of the CoNLL-X '06*, New York City, New York.

Allan Bay. 2003. *Cuochi si diventa. Le ricette e i trucchi della buona cucina italiana di oggi*, volume 1. Feltrinelli.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *Web Semant.*, 7(3):154–165, September.

Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Leonardo Lesmo. 200. Building a treebank for italian: a data-driven annotation schema. In *In Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000*, pages 99–105.

Massimiliano Ciaramita. 2003. Supersense tagging of unknown nouns in wordnet. In *In EMNLP 2003*, pages 168–175.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

C. D'Onofrio. 1997. *Il Cucchiaio d'Argento*. Editoriale Domus. On-line version: http://www.cucchiaio.it/ricette/ricetta-caponata-classica .

Gregory Druck. 2013. Recipe Attribute Prediction using Review Text as Supervision. In *Cooking with Computers 2013, IJCAI workshop*.

FAO. 2014. AGROVOC Project. http://aims.fao.org/standards/agrovoc/.

Christiane Fellbaum. 2005. Wordnet and wordnets. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford. Elsevier.

M. F. K. Fisher, 2001. *The Anatomy of a Recipe*, chapter 1, pages 13–24. Vintage.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, January.

Haoran Xie and Lijuan Yu and Qing Li. 2010. A Hybrid Semantic Item Model for Recipe Search by Example. In *ISM*, pages 254–259.

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

D. Jurafsky. 2014. *The Language of Food: A Linguist Reads the Menu*. W. W. Norton.

Leonardo Lesmo, Alessandro Mazzei, Monica Palmirani, and Daniele P. Radicioni. 2013. TULSI: an NLP system for extracting legal modificatory provisions. *Artif. Intell. Law*, 21(2):139–172.

Leonardo Lesmo. 2007. The rule-based parser of the NLP group of the University of Torino. *Intelligenza artificiale*, 2:46–47.

Hirokuni Maeta, Shinsuke Mori, and Tetsuro Sasada. 2014. A framework for recipe text interpretation. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 553–558, New York, NY, USA. ACM.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Yuka Shidochi, Tomokazu Takahashi, Ichiro Ide, and Hiroshi Murase. 2009. Finding replaceable materials in cooking recipe texts considering characteristic cooking actions. In *Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities*, CEA '09, pages 9–14, New York, NY, USA. ACM.

Kirstin Walter, Mirjam Minor, and Ralph Bergmann. 2011. Workflow extraction from cooking recipes. In Belen Diaz-Agudo and Amelie Cordier, editors, *Proceedings of the ICCBR 2011 Workshops*, pages 207–216.

Wikibooks. 2014. Wikibooks, manuali e libri di testo liberi: Libro di cucina - Ricette. it.wikibooks.org/wiki/Libro_di_cucina/Ricette/Caponata .