Dish Classification using Knowledge based Dietary Conflict Detection

Nadia Clairet
LIRMM (Laboratoire d’Informatique, de Robotique et de Microélectronique de Montpellier)
860 rue de St Priest 34000 Montpellier, France
LIMICS (Laboratoire d’Informatique Médicale et d’Ingénierie des Connaissances en e Sant)
74 rue Marcel Cachin 93017 Bobigny, France
Lingua et Machina
7 Boulevard Anatole France, 92100 Boulogne-Billancourt, France
clairet@lirmm.fr

Abstract
The present paper considers the problem of dietary conflict detection from dish titles. The proposed method explores the semantics associated with the dish title in order to discover a certain or possible incompatibility of a particular dish with a particular diet. Dish titles are parts of the elusive and metaphoric gastronomy language, their processing can be viewed as a combination of short text and domain-specific texts analysis. We build our algorithm on the basis of a common knowledge lexical semantic network and show how such network can be used for domain specific short text processing.

1 Introduction
The dish classification according to some dietary criteria is a challenging task. Performing such classification given the dish title as it may appear in a restaurant menu implies solving the following issues: context scarcity, relevant information selection, and output qualification. In the framework of our method, these issues have been dealt with following two steps. First, domain specific knowledge has been immersed into a large general knowledge lexical semantic network. Second, a graph traversal algorithm has been defined and parametrized in order to grasp relevant lexical and semantic information for the dietary conflict detection.

In terms of structure, the network we use is a directed graph where nodes may represent simple or compound terms, linguistic information, phrasal expressions, and sense refinements. The arcs of the graph are directed, weighted, and typed according to the ontological, semantic, lexical associations between the nodes. They also may be semantically annotated (ex.: quiche part−whole frequent egg) which is convenient for working with domain specific expert knowledge. During the traversal (interpretation) of the graph, the nodes and the arcs are referred to as respectively terms and relationships. Throughout this paper, relationship corresponds to a quadruplet \( R = \{\text{term}_{\text{source}}, \text{type}, \text{weight}, \text{term}_{\text{target}}\} \). Weight denotes an association force of the relationship between two terms of the graph.

The paper will be structured as follows. First, we will recall main state-of-the-art achievements related to short text analysis and cooking recipe analysis. Second, we will detail materials of the approach: knowledge resource structure, text pre-processing steps. Third, we will describe our graph traversal method. Finally, we will present and discuss the output of our system and introduce its possible evolutions.

2 State of the Art

2.1 Distributional Methods for Short Text Analysis
The processing of dish titles can be considered as a possible specialization of some general method for short text analysis. Given the importance of the distributional hypothesis in the current NLP practice, we will empathize knowledge based distributional method and logic based distributional method to the short text processing. A relevant knowledge based distributional method has been proposed by (Hua et al., 2015) in the framework of the Probase lexical semantic network. The

---

1Terms and relationships related to nutrition, cooking techniques, cooking recipes and their typical parts, diets and typical dietary restrictions.
2Unlike the dictionary sense, the sense refinement reflects the typical use of a term, its sense activated by a particular context.

---

The main idea behind the distributional hypothesis: "a word is characterized by the company it keeps" (Firth, 1957).
Probase associated tools allow obtaining a concept distribution based on different scoring functions (conditional probability distribution\(^3\), point-wise mutual information\(^6\) etc.). Prior to the analysis of short texts, the authors in (Hua et al., 2015) acquire knowledge from web corpus, Probase network as well as a verb and adjective dictionary. They solve the segmentation task by introducing a multi-fold heuristic for simple and multi-word term detection. It takes into account the presence of *is-a* and co-occurrence relationships between the candidate terms. Subsequently, the terms are typed according to different categories: part of speech, concept/entity distinction etc. Finally, the disambiguation is done using the weighted vote (conceptual connections of the candidate term considering the "vote of context"). This method seems to be relevant for queries, however it would be difficult to use it for the dish classification. The main difficulty comes from the fact that it is concept driven. Indeed, for a term such as "creme brulee", we obtain the concept distribution scores shown in Table 1. Due to the underlying semantic relationship types (*is-a*, *relatedness*), these examples bring very few information about the composition and the semantic environment (that needs considering relationship types expressing *part-whole*, *location*, *instrument*, *sensory characteristic* relationships) of the dish titles and one can hardly qualify some of the returned scores (e.g. *off beat flavor*) in order to approximate the underlying relationship type. An improvement of the knowledge based distributional method could be made using an additional semantic resource containing a rich part-whole semantics\(^7\) such as WordNet (Fellbaum, 1998).

A different kind of methods for short text analysis relies on general-purpose first order probabilistic logic as shows the hybrid approach developed by (Beltagy et al., 2014). The authors use them to generate an on-the-fly ontology that contains the relevant information related to some current semantic analysis task. They adopt probabilistic logic as it allows weighted first order logic formulas. The weight of the formulas corresponds to a certainty measure estimated from the distributional semantics (meaning representation, semantic relation prediction). First, natural language sentences are mapped to a logical form (using the Boxer tool (Bos, 2015)\(^8\)). Second, the ontology is encoded in the form of weighted inference rules describing the semantic relations (the authors mention *hyponymy*, *synonymy*, *antonymy*, *contextonymy* i.e. the relation between "hospital" and "doctor"). Third, a probabilistic logic program answering the target task is created. Such program contains the evidence set, the rule base (RB, weighted first order logical expressions), and a query. It calculates the conditional probability \(P(\text{Query} | \text{Evidence}, \text{RB})\). This approach has been tested on the SICK\(^9\) data for the tasks of semantic textual similarity detection and textual entailment recognizing. The results showed a good performance of this method over distributional only and logic only methods. This kind of approaches rather considers lexical relations (revealing some linguistic phenomena), than purely semantic (language independent, world knowledge phenomena) relations. As the knowledge based distributional method, it suffers from the difficulty to qualify the relationships as it uses mainly the co-occurrence analysis.

| Term (concept)               | MI   | \(P(\text{concept}|\text{entity})\) |
|-----------------------------|------|-------------------------------------|
| dessert                     | 0.386| 0.446                               |
| authentic bistro dish       | 0.164| 0.12                                |
| off beat flavor             | 0.047| 0.06                                |
| homemade dessert            | 0.046| 0.048                               |
| dairy based dessert         | 0.04  | 0.036                               |

Table 1: Concept distribution for *creme brulee* (Probase conceptualization tool)

---

\(^3\)Conditional probability distribution of two discrete variables (words)

\[ P(A|B) = \frac{P(A) \cap (B)}{P(B)} \]

\(^6\)Point-wise Mutual Information is calculated according to the formula:

\[ PMI(x;y) = \log \frac{P(x,y)}{P(x)P(y)} \]

\(^7\)The WordNet part-whole relation splits into three more specific relationships: *member-of, stuff-of, and part-of.*

\(^8\)Boxer (Bos, 2015) is a semantic parser for English texts based on Discourse Representation Theory.

\(^9\)Sentences Involving Compositional Knowledge. This dataset includes a large number of manually annotated sentence pairs that are rich in the lexical, syntactic and semantic phenomena (e.g., contextual synonymy, syntactic alternations). [http://clic.cimec.unitn.it/composes/sick.html](http://clic.cimec.unitn.it/composes/sick.html).
2.2 Cooking Instruction Analysis

Besides the approaches focused on flavor networks which consider cooking recipes as "bags of ingredients" such as (Ahn et al., 2011), the existing approaches to the recipe analysis concentrate on recipes taken as a sequence of instructions that can be mapped to a series of actions. Numerous publications concern the implementation of supervised learning methods. For instance, (Mori et al., 2012) use the annotated data to extract predicate-argument structures from cooking instructions in Japanese in order to represent the recipe as a work flow. The first steps of this process are words segmentation and entity type recognition based on the following entity types: Food, Quantity, Tool, Duration, State, chefs action, and foods action. Therefore, this task is similar to the conceptualization process proposed by (Hua et al., 2015) discussed in the previous section. Entity type recognition is followed by syntactic analysis that outputs a dependency tree. The final step aims at extracting predicate-argument triples from the disambiguated (through segmentation, entity type recognition, and dependency parsing) recipe text. In this approach, the semantic information that could be attached to the arcs is attached to the nodes. The node type together with the syntactic markers (i.e. case marker) helps determining the nature of the predicate-argument relation. This method yields modest results and could be improved by adopting a more versatile graph structure (i.e. a structure with typed arcs). (Kiddon et al., 2015) proposed a similar unsupervised technique for mapping recipe instructions to actions based on a hard Expectation Maximization algorithm and a restricted set of verb argument types (location, object).

In the paradigm of semantic role labeling, the approach of (Malmaud et al., 2014) uses a Markov decision process where ingredients and utensils are propagated over the temporal order of instructions and where the context information is stored in a latent vector which disambiguates and augments the instruction statement under analysis. The context information corresponds to the state of the kitchen and integrates the changes of this state according to the evolving recipe instructions. It yields a good performance for the instruction analysis.

In the case based reasoning paradigm, (Dufour-Lussier et al., 2012) represent the cooking instructions as a work-flow and thus propose a method for the automatic acquisition of a rich case representation of cooking recipes for process-oriented case-based reasoning from free recipe text. The cooking process is represented using the Allen (Allen, 1981) algebra extended with relations over interval durations. After applying classical NLP tools for segmentation, part-of speech tagging, syntactic analysis, the extraction process from texts focuses on the anaphora resolution and verb argument analysis. The actions are modeled on the basis of the instructional text without considering the implicit information proper to the cooking recipes. The underlying knowledge resource is the case-based reasoning system Taaable (Cordier et al., 2014). The existing methods of cuisine texts analysis converge on the necessity to have a richer context around the terms present in the input text. Various approaches have been proposed to provide such context. Some of them propose specific meta-languages such as MILK\textsuperscript{10} (Tasse and Smith, 2008) or SIMMR (Jermsurawong and Habash, 2015)\textsuperscript{11}. MILK has been proposed as a machine-readable target language to create sets of instructions that represent the actions demanded by the recipe statements. It is based on the first-order logic, but allows handling the temporal order as well as creation/deletion of ingredients. A small corpus of 250 recipes has been manually annotated using MILK (CURD (Carnegie Mellon University Recipe Database)). Similarly, SIMMR (Jermsurawong and Habash, 2015) allows to represent a recipe as a dependency tree. The leaves of the tree are the recipe ingredients and its internal nodes are the recipe instructions. MILK tags have been used to construct the SIMMR trees. Machine learning methods (SVM\textsuperscript{12} classification) have then been used for instruction-ingredient linking, instruction-instruction linking using SIMMR. Other approaches for getting more context for the cooking recipe analysis are dynamic data structures (i.e. latent vector used by (Malmaud et al., 2014)) and graph-shaped resources (ontologies, semantic networks) as in our case.

\textsuperscript{10}Minimal Instruction Language for the Kitchen.
\textsuperscript{11}Simplified Ingredient Merging Map in Recipes.
\textsuperscript{12}Support Vector Machine Classification.
3 Preliminaries

The dietary restrictions are defined as follows: Diabetes, GlutenFree, Halal, Hindu, Kosher, LowCalories, LowFat, LactoseFree, LowSalt, Vegan, Vegetarian. For these diets, we automatically extracted and manually validated an initial set of approximately 200 forbidden ingredients. We used domain-specific resources (lists of ingredients) found on the Web for this step. These ingredients (if not already present in our graph-based knowledge resource) have been encoded as nodes and linked to the nodes representing the diets listed above by the arcs typed $r_{incompatible}$. The nutritional restrictions differ in terms of their semantic structure (which may rely on composition (part-whole relation), cutting types (holonymy etc.), cooking state (i.e. boiled carrots are undesirable in case of diabetes).

3.1 Corpus and Knowledge Resource

For our experiment we used a set of 5,000 dish titles in French, which corresponds to a corpus of 19,000 words and a vocabulary of 2,900 terms (after removing stop words and irrelevant expressions such as "tarte tatin à ma façon", tarte tatin my way). This data mirrors the French urban culinary tradition as well as the established practice of searching the Web for recipes. The data scarcity proper to the recipe titles is an important obstacle to their processing. Distributional scores used by (Hua et al., 2015) and (Beltagy et al., 2014) may be interesting in terms of flavor associations, but they also demonstrate that we need to know more about the most likely semantic neighborhood of the term to handle the part-whole semantics and incompatibility analysis.

The semantic resource we use for the experiments is the RezoJDM lexical semantic network for French. This resource stems from the game JeuxDeMots (contribution interface for RezoJDM) played partly through the direct contribution. External resources such as domain specific lexicons and terminological resources have been used. In particular, some equivalence and synonym relation triples have been extracted from IATE term base. The Agrovoc thesaurus provided only a few new terms; it contained no relevant relations. Additionally, as RezoJDM can be enhanced using crowd-sourcing methods and, in particular, games with a purpose, specific game with a purpose assignments have been given to the JeuxDeMots (contribution interface for RezoJDM) players. Today (June 2017) the domain specific subgraph corresponds to 40K terms (approximately 2.8% of the RezoJDM). The overall adaptation process took about 3 weeks.

In the framework of our experiment, we use taxonomy relations, part-whole relations, object-mater relations and, in some cases, hyponymy and characteristic relations. Running the experiment involved preprocessing steps (multi-word terms detection, lemmatization, disambiguation), browsing

---

13The corpus has been collected from the following Web resources: 15% www.cuisineaz.fr, 20% www.cuisinellibre.org et 65% www.allrecipe.fr
14http://www.jeuxdemots.org/jdm-about.php
15http://iate.europa.eu/switchLang.do?success=mainPagelang=fr
16The part of the IATE resource that has been used corresponds to the subject field 6006.
17http://aims.fao.org/fr/agrovoc
the graph following path constraints, scoring possible incompatibilities, and possibly normalizing scores. Among other approaches using a similar plot, (Poria et al., 2014) use ConceptNet (Liu and Singh, 2004) network for the semantic analysis task.

4 Method

The preprocessing step includes text segmentation which relies on the multi-word lexical entities detection, followed by stop-words removal. The multi-word term detection is done in two steps. First, the recipe titles are cut into n-grams using a dynamic-length sliding window (2 ≤ 4). Second, the segments are compared to the lexical entries present in RezoJDM which is therefore used as a dictionary. In RezoJDM, the multi-word terms are related to their "parts" by the relation typed location (or "phrase"). This relation is useful for compound terms analysis. Bi-grams and trigrams are the most frequent structures in the domain-specific subgraph as well as in RezoJDM. Indeed, they represent respectively 28% and 16% of the overall set of multi-word terms and expressions (300K units). For our experiment, we often opted for trigrams which turned out to be more informative (i.e. correspond to nodes with a higher out degree).

| French (English) |
|-----------------|
| 1. Carpe à la forestière (Forest style carp) |
| 2. Carré d’agneau dans son jus (Rack of lamb in its own jus) |
| 3. Waterzooi de lieu noir (Coley waterzooi) |
| 4. Verrines de lentilles au saumon fumé (Lentil and smoked salmon verrines) |
| 5. Gâteau basque à la confiture de cerises noires (Basque cake with black cherry jam) |

Table 2: Preprocessed recipe titles. Expressions relevant for the domain (1,2) as well as multi-word terms (3,4), and phrases(5) are detected. Sometimes, two segmentations are possible for the same input. In such cases, the contrast is favored by exploring syntactic features and outgoing semantic relationships of terms in RezoJDM.

4.1 From Text to Graph

Starting from a sequence w₁, w₂, ..., wₙ of n terms, we build a context namely the lemmatized and disambiguated representation of the recipe title under scope. Such context is the entry point to our knowledge resource.

A context is a sequence of nodes (the text order is preserved): Cₜₜₐₜ₁, w₂, ..., wₙ = n₁, n₂, ..., nₙₚₙ where the node nᵢ is the most precise syntactic and semantic representation of the surface form available in RezoJDM. To obtain such representation we search for the node corresponding to the surface form if it exists. Then, we yield its refinement if the term is polysemic. The irrelevant refinements are discriminated using a list of key terms that define our domain (i.e. thematic subgraph within a lexical semantic network). The identification of the refinement is done through the cascade processing of relationships of a node typed refinement, domain, and meaning. The context creation function returns for quiche au thon et aux tomates (quiche with tuna and tomatos) the result {quiche (préparation culinaire), thon (poisson, chair), tomate légume-fruit], respectively: “quiche (preparation), tuna (fish flesh), tomato (fruit-vegetable)"}

For each node nᵢ ∈ C we explore paths S = ((n₁a₁n₂), (n₂a₂n₃), (nₙ₋₁aₙ₋₁aₙ)). The type of relationships we choose for the graph traversal depends on the local category of the node:

1. If isa("preparation", nᵢ), r ∈ {hypo, part – whole, matter}. This is the case of mixtures, dishes and other complex ingredients;
2. If isa("ingredient", nᵢ), r ∈ {isa, syn, part – whole, matter, charac}. It is the case of plain ingredients like tomato.

The weight of all relations we traverse must be strictly positive. We traverse the graph testing a range of conditions: relevance to the domain interest Dₐₐₐₐₐₐₐ (food domain), existence of a path of a certain length (≤ 2) and type between the candidate node and and the rest of the Context under analysis, co-meronymy relation etc.

To obtain relevant results, two conditions are to be fulfilled. First, there must be a disambiguation strategy and a domain filtering. Second, the similarity has to be handled between the preparation and its hyponyms. Indeed, the exploration of the part-whole relations of the network refers to all the possible ingredients and constituents

---

18 In the opposite case the acquisition process through external resources like Wikipedia, Web and through crowdsourcing (Games with a purpose) may be triggered (if an open world hypothesis is favored).
19 First Order Logic Notation.
20 Idem.
of the preparation. If the preparation has a conceptual role (i.e. "cake"), the part-whole relations analysis will output a lot of noise. In our example, it is important to grasp the absence of pork (meat) and the presence of tuna (fish) in the quiche under scope. Therefore, instead of directly exploring the part-whole relations of the quiche(preparation), we rather try to find similar quiche(preparation) hyponyms for the context $C$="quiche (preparation), tuna (fish flesh), tomato (fruit-vegetable)" and yield the typical parts they have in common with the generic quiche(preparation). Our function finds the hyponym which maximizes the similarity score. This score is a normalized Jaccard index over all the positive outgoing part-whole, isa, matter relations.

$$J(S_C, S_{Chypo}) = \frac{s_{C\cap S_{Chypo}}}{s_{C\cup S_{Chypo}}}$$

where $S_{Chypo}$ is the hyponym context built on the go.

Different threshold values have been experimented. Empirically, the threshold fixed at 0.30 allows capturing generic similarities such as (for our example) quiche saumon courgette ("quiche salmon zucchini") score=0.32, a quiche with some fish and a vegetable. More precise similarity corresponds to higher scores.

Using the described strategy, the irrelevant recipes such as quiche lorraine (score=0.20) are efficiently discriminated. Once the relevant hyponym is identified, its part-whole neighbors can be grasped. A specific graph traversal strategy is used for the LowSalt diet. It includes exploring the characteristic relation type for the preparation and its parts.

4.2 From Graph to Incompatibilities

The incompatibility calculation takes as input the list of diets and the queue $F$ containing terms related to main context. This function looks for an dietary incompatibility path $S_{inc}$ such that $N' \in F$ and $S_{inc} = (\langle N', r_{type}, N \rangle, (N, r_{inc}, N_{Diet}))$ and $type \in \{ holo\|isa\|haspart\|substance\|hypo \}$.

The output is a key value pair (diet, score). The score depends on the distance in the RezoJDM graph and on the relation type between the diet and the incompatible node in the context. Besides the refinement relation type\(^{21}\), it is calculated as follows for the distance $d := \frac{1}{1+d}$. The score for the whole context corresponds to the addition of the individual scores of the context nodes. It is adapted in order to bring it closer to a probability distribution and allow further statistical or predictive processing. Our system first obtains a list of nodes linked to the context (see 3.1). After following a traversal strategy, it returns a list of probabilistic scores for each part of the context. I.e. ($w$ corresponds to weight and $d$ corresponds to distance):

$$\text{recipe : quiche thon tomate}$$

$$\text{quiche>preparation,thon>poisson,}$$

$$\text{tomate>legume}$$

Further follows similarity processing as described in 3.2., then processing of parts shared by quiche and its hyponyms similar to the context. For each part, the isa, part-whole, mater and characteristic relations are further explored

$$r\_has\_part\ pte\ d=1\ w=6\ *eggs*\ etc.$$

Intermediate raw scores are returned (example : quiche(preparation)):

| Diabetes 1.3, LactoseFree 3.0, Halal 0.0, Kosher 0.3, LowCalories 1.5, LowSalt 0.5, GlutenFree 1.8, Hindu 0.5, LowFat 0.8, Vegan 0.5, Vegetarian 0.3 |

Same processing is performed for all the other nodes from the context. The output summarizes scores per ingredient and per diet. It can be "normalized" according to the following rule applied to the raw score in order to $s \in L_5$ (raw list of scores) in order to produce a probabilistic score $s_p$ : if $s \geq 0.5, s_p \leftarrow 1$, if $s \leq 0.5, s_p \leftarrow 0.5$.

The range of this new score is restricted to three values : compatible (0), uncertain (0.5), and incompatible (1).

$$\text{final probabilistic scores for *quiche thon tomate*}$$

| Diabetes 0.5, LactoseFree 1.0, Halal 0.0, Kosher 0.5, LowCalories 1.0, LowSalt 1.0, GlutenFree 1.0, Hindu 1.0, LowFat 1.0, Vegan 1.0, Vegetarian 1.0 |

In a restaurant scenario, a client would be informed about the strict incompatibility or

\(^{21}\)Which participates to the disambiguation process.
compatibility of a dish with his or her nutritional restrictions and alerted about some potential incompatibilities that would need further information from the caterer. The list of terms that are not present in the resource (for example, *sibneh, zwiebelkuchen*) is output by the system. It serves to the further improvement of the RezoJDM graph from external resources.

### 4.3 Evaluation and Discussion

The system has been evaluated using an incompatibility annotated corpus of 1,500 recipe titles. The evaluation data has been partially collected using structured document retrieval simultaneously with the raw corpus constitution. Relevant metatags (HTML tags present in micro-formats) have been used to pre-build incompatibility scores. The overlap between the different diets (vegan recipes are compatible with a vegetarian diet etc.) has been taken into account. The labels have been adjusted and enriched (LawSalt, Kosher diets) by hand by the author because the structured metatags do not contain such information. The obtained evaluation scores are binary. Our system is under improvement, we estimated that the score returned by the system should be $\geq 0.5$ to be ranked as acceptable. Later, a finer grained evaluation will be adopted.

Our results expressed in terms of precision, recall and f-measure (F1 score) are listed in Table 4. The most important score is the precision as, for an allergic or intolerant restaurant customer, even a very small quantity of a forbidden product may be dangerous. The average value correspond to the macro-average (arithmetic mean), F-score average is therefore the harmonic mean of precision average rate and recall average rate.

For the halal diet, there are very few terms in the graph that point to this diet. The *low calories* and *vegetarian* diets are well known among the RezoJDM community and well represented in the graph. In 13% of cases, the graph traversal did not return any result as the terms corresponding to the words of the context do not exist in the graph. It was the case of borrowings (such as *quinotto*) or lexical creations (i.e. *anti-tiramisu* as our analysis scheme doesn’t take into account morphology). The average number of incompatibility scores $\neq 0$

---

22 Micro-formats($\mu F$) refer to standardized semantic markup of web-pages.

23 The term *lexical creation* refers to the neology (creation of new terms).

---

| Diet            | Precision | Recall | F1 score |
|-----------------|-----------|--------|----------|
| Diabetes        | 92%       | 92%    | 92%      |
| LactoseFree     | 71%       | 73%    | 72%      |
| Halal           | 65%       | 75%    | 70%      |
| Kosher          | 67%       | 60%    | 63%      |
| LowCalories     | 60%       | 75%    | 67%      |
| LowSalt         | 88%       | 65%    | 75%      |
| GlutenFree      | 80%       | 73%    | 76%      |
| Hindu           | 86%       | 80%    | 83%      |
| LowFat          | 67%       | 70%    | 68%      |
| Vegan           | 80%       | 90%    | 85%      |
| Vegetarian      | 83%       | 90%    | 86%      |
| **macro-average** | **76%**  | **77%** | **76%** |

Table 4: Corpus-based evaluation. We specify the repartition between the 3 possible probabilistic scores in the whole corpus.
The mean \((M)\) and standard deviation \((SD)\) values reveal this confidence issue (Table 5). A low \(M\) value together with a low \(SD\) indicates that there were too many uncertain scores \((\leq 0.5)\) among the resulting scores of a particular diet. On the contrary, \(M \geq 0.5\) and \(SD \leq 0.5\) reveal a good confidence about the incompatibility. In the case of ”no salt” diet, very few incompatibilities have been detected with low confidence, thus finer grained approach is needed (i.e. relationship annotation). In the case of GlutenFree and LowCalories the scores are quite confident but sparse : the data is unequally distributed in the resource, more relationship types should be explored. The overall \(M\) and \(SD\) values reveal low confidence but uniform scores. To maximize the confidence related to the score, the knowledge resource must be continuously enhanced. Today this is achieved using two main strategies: exogenous and endogenous. The first one may include term extraction from corpora, relationship identification based on word embeddings, direct contribution, specific games with a purpose assignments. The second one refers to propagation of the relations already existing in the graph using inference schemes based on the transitivity of the \(isa\), \(part-whole\), and \(synonym\) semantic relations as proposed by (Zarrouk et al., 2013). Using a lexical semantic network for text processing offers some clear advantages. First, there is no need to use multiple data structures during the processing. Second, the graph structure supports encoding various kinds of information that can be useful for text processing tasks. Every piece of information encoded using the graph formalism is machine interpretable and can be accessed during the traversal. Finally, the graph based analysis strategy has an explanatory feature, it is always possible to know why the system returned some particular output. However, two main issues must be tackled : the knowledge resource population method and the filtering strategy while traversing the resource. To test the portability of the approach, we run it using the ConceptNet common knowledge network and a restricted set of aligned recipe titles in English and French. The obtained scores have been compared across languages. They are given as follows in our example: ”score for French [score for English]".

| Diet          | \(M\)  | \(SD\)  |
|---------------|-------|-------|
| Diabetes      | 0.785 | 0.295 |
| LactoseFree   | 0.200 | 0.215 |
| Halal         | 0.240 | 0.239 |
| Kosher        | 0.275 | 0.243 |
| LowCalories   | 0.487 | 0.484 |
| LowSalt       | 0.031 | 0.094 |
| GlutenFree    | 0.549 | 0.349 |
| Hindu         | 0.447 | 0.244 |
| LowFat        | 0.380 | 0.202 |
| Vegan         | 0.406 | 0.219 |
| Vegetarian    | 0.809 | 0.253 |
| overall       | 0.406 | 0.219 |

Table 5: Mean values per diet (evaluation)

Thus, the portability of the method is proportional to the quantity of available resources that can be assimilated to human contribution (i.e.corpora of user contributed recipes for a given language) and to the amount of direct human expert or non expert contribution.

5 Conclusion

We introduced the use of a lexical-semantic network for dietary conflict detection based on dish titles. The scores obtained by our system can be used for building specific resources for machine learning tasks such as classifier training. Among the perspectives, we can list :

- knowledge resource population using external resources and endogenous approaches ;
- mapping the resource to other existing semantic resources;
- making the system evolve towards a multilingual (language independent?) dietary conflict detection.

An important advantage of the system over purely statistical approaches is its explanatory feature, we can always know how the system came to its decision and thus can constantly improve it. The limitations of our contribution are linked to its improvement model which is based on contribution work and the necessity to (weakly) validate the new relations.
References

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, and Albert-Laszl Barabsi. 2011. Flavor network and the principles of food pairing. *CoRR* abs/1111.6074.

James F. Allen. 1981. An interval-based representation of temporal knowledge. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI ’81, Vancouver, BC, Canada, August 24-28, 1981*. pages 221–226.

I. Beltagy, Katrin Erk, and Raymond Mooney. 2014. Semantic parsing using distributional semantics and probabilistic logic. In *Proceedings of ACL 2014 Workshop on Semantic Parsing (SP-2014)*. Baltimore, MD, pages 7–11.

Johan Bos. 2015. Open-domain semantic parsing with boxer. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Institute of the Lithuanian Language, Vilnius, Lithuania*. pages 301–304.

Stergios Chatzikyriakidis, Mathieu Lafourcade, Lionel Ramadier, and Manel Zarrouk. 2015. Type Theories and Lexical Networks: Using Serious Games as the Basis for Multi-Sorted Typed Systems. In *ESSLLI: European Summer School in Logic, Language and Information*. Barcelona, Spain.

Amélie Cordier, Valmi Dufour-Lussier, Jean Lieber, Emmanuel Nauer, Fadi Badra, Julien Cojan, Emmanuelle Gaillard, Laura Infante-Blanco, Pascal Molli, Amedeo Napoli, and Hala Skaf-Molli. 2014. Taxable: a Case-Based System for personalized Cooking. In Stefania Montani and Lakhmi C. Jain, editors, *Successful Case-based Reasoning Applications-2*, Springer, volume 494 of *Studies in Computational Intelligence*, pages 121–162.

Valmi Dufour-Lussier, Florence Le Ber, Jean Lieber, Thomas Meilender, and Emmanuel Nauer. 2012. Semi-automatic annotation process for procedural texts: An application on cooking recipes. *CoRR* abs/1209.5663.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-55. 1952-59:1–32.

Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2015. Short text understanding through lexical-semantic analysis. In Johannes Gehrke, Wolfgang Lehner, Kyuseok Shim, Sang Kyun Cha, and Guy M. Lohman, editors, *ICDE*. IEEE Computer Society, pages 495–506.

Jermsak Jerm surawong and Nizar Habash. 2015. Predicting the structure of cooking recipes. In Lluis Mquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *EMNLP*. The Association for Computational Linguistics, pages 781–786.

Chlo Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. *Mise en place: Unsupervised interpretation of instructional recipes*, Association for Computational Linguistics (ACL), pages 982–992.

Mathieu Lafourcade. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP’07: 7th International Symposium on Natural Language Processing*, Pattaya, Chonburi, Thailand, page 7.

Mathieu Lafourcade. 2011. *Lexicon and semantic analysis of texts - structures, acquisition, computation and games with words*. Habilitation à diriger des recherches, Université Montpellier II - Sciences et Techniques du Languedoc.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal* 22(4):211–226.

Jonathan Malmaud, Earl Wagner, Nancy Chang, and Kevin Murphy. 2014. Cooking with semantics. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*. Association for Computational Linguistics, Baltimore, MD, pages 33–38.

Shinsuke Mori, Tetsuro Sasada, Yoko Yamakata, and Koichiros Yoshino. 2012. A machine learning approach to recipe text processing.

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.

Soujanya Poria, Basant Agarwal, Alexander Gelbukh, Amir Hussain, and Newton Howard. 2014. *Dependency-Based Semantic Parsing for Concept-Level Text Analysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 113–127.

Lionel Ramadier. 2016. *Indexing and learning of terms and relations from reports of radiology*. Theses, Université de Montpellier.

Dan Tasse and Noah A. Smith. 2008. Sour cream:toward semantic processing of recipes. *T.R. CMU-LTI-08-005* page 9.

Manel Zarrouk, Mathieu Lafourcade, and Alain Joubert. 2013. Inference and Reconciliation in a Crowdsourced Lexical-Semantic Network. In *CICLING: International Conference on Intelligent Text Processing and Computational Linguistics*. Samos, Greece, 14th.