Automatic identification of domain terms: an approach for Italian

Maria Teresa Artese, Isabella Gagliardi
IMATI – CNR
Via Bassini 15, 20133 Milan, Italy
{artese,gagliardi}@mi.imati.cnr.it

DiPP 2020 : Digital Presentation and Preservation of Cultural and Scientific Heritage

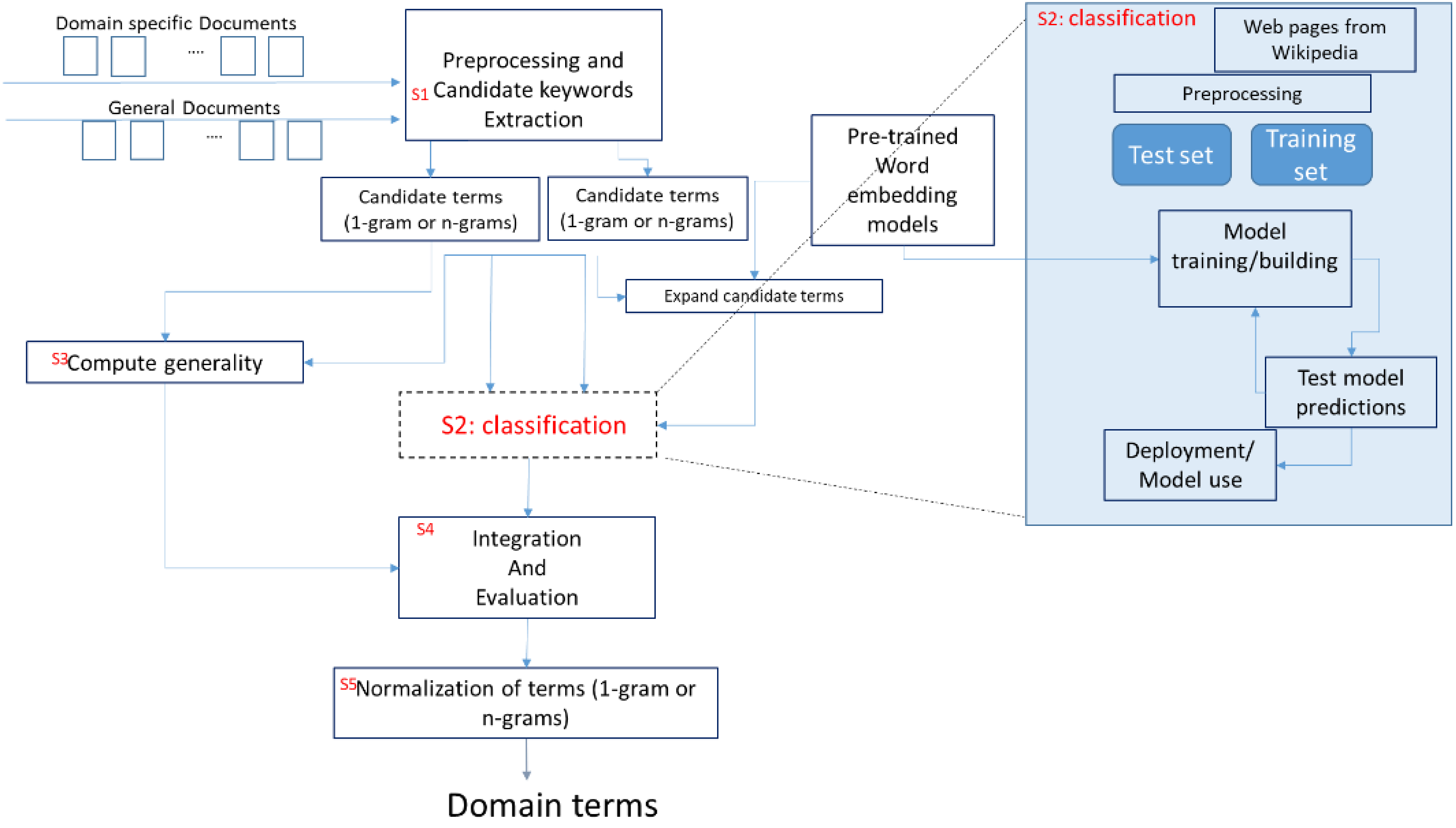
Characteristics of the approach

- Integration of different methods:
  - ML classification methods working on
  - Word embedding models, the semantic representation of candidate terms

  And
  - a computation of the degree of specialization of a term (called generality)

- Easy to implement and apply to other domains

- Work in progress



Domain specific Documents

....

General Documents

....

S1 Preprocessing and Candidate keywords Extraction

Candidate terms (1-gram or n-grams)

Candidate terms (1-gram or n-grams)

Pre-trained Word embedding models

Expand candidate terms

S3 Compute generality

S2: classification

S4 Integration And Evaluation

S5 Normalization of terms (1-gram or n-grams)

Domain terms

S2: classification

Web pages from Wikipedia

Preprocessing

Test set

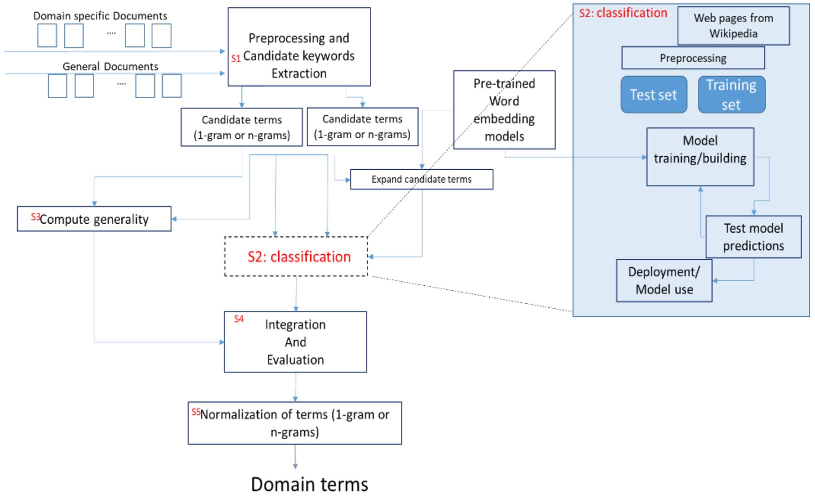
Training set

Model training/building

Test model predictions

Deployment/ Model use

Steps of the approach



Domain specific Documents

....

General Documents

....

S1 Preprocessing and Candidate keywords Extraction

Candidate terms (1-gram or n-grams)

Candidate terms (1-gram or n-grams)

Pre-trained Word embedding models

Expand candidate terms

S3 Compute generality

S2: classification

S4 Integration And Evaluation

S5 Normalization of terms (1-gram or n-grams)

Domain terms

S2: classification

Web pages from Wikipedia

Preprocessing

Test set

Training set

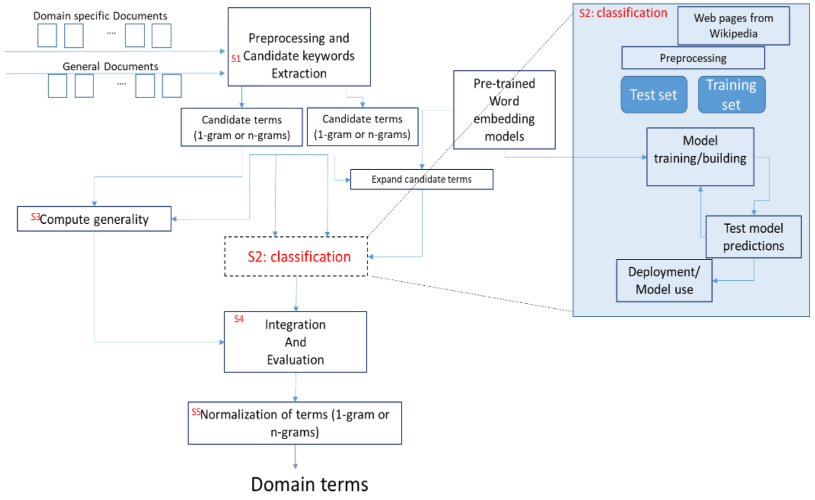
Model training/building

Test model predictions

Deployment/ Model use

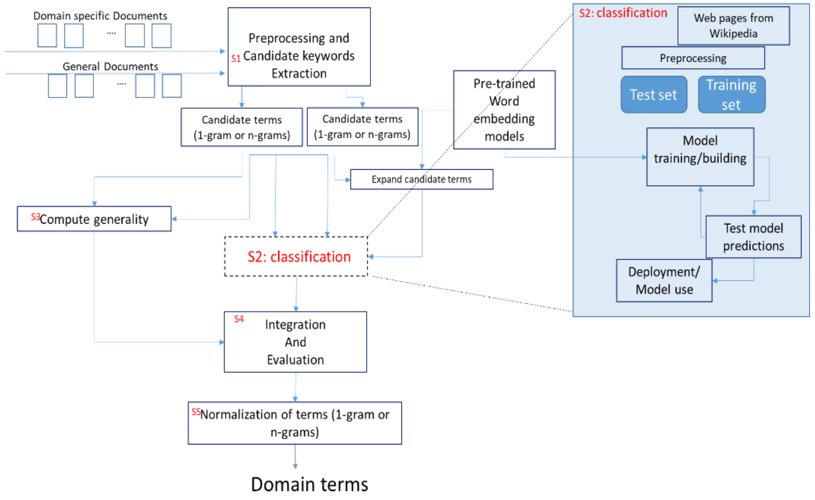
- Expand candidate terms: as our aim here is to define a set of domain terms, we are interested in enlarging as much as possible the candidate terms to be classified as belonging to the domain of interest. For each original candidate term, the n most similar terms, according to the word embedding model adopted, were classified too.

Steps of the approach


Domain specific Documents
....
General Documents
....
S1 Preprocessing and Candidate keywords Extraction
Candidate terms (1-gram or n-grams)
Candidate terms (1-gram or n-grams)
Pre-trained Word embedding models
Expand candidate terms
S3 Compute generality
S2: classification
S4 Integration And Evaluation
S5 Normalization of terms (1-gram or n-grams)
Domain terms
S2: classification
Web pages from Wikipedia
Preprocessing
Test set
Training set
Model training/building
Test model predictions
Deployment/ Model use

- **S2: classify candidate terms**,
  as belonging /not belonging to a specific domain.
- *similarity/distance measure* is needed to evaluate the belonging of observations to a category from a set of categories defined ahead, based on the training set of data whose category is known.
- The method used here to calculate relatedness among words/sentences is based *on word embeddings*.
  Each word is represented as a real value vector in a predefined vector space
- *Word2Vec* is one of the most used techniques to learn word embedding using shallow neural networks
- The Global Vectors for Word Representation, or *GloVe* algorithm is an extension to the Word2Vec method for efficiently learning word vectors.

Steps of the approach


Domain specific Documents
....
General Documents
....
S1 Preprocessing and Candidate keywords Extraction
Candidate terms (1-gram or n-grams)
Candidate terms (1-gram or n-grams)
Pre-trained Word embedding models
Expand candidate terms
S3 Compute generality
S2: classification
S4 Integration And Evaluation
S5 Normalization of terms (1-gram or n-grams)
Domain terms
S2: classification
Web pages from Wikipedia
Preprocessing
Test set
Training set
Model training/building
Test model predictions
Deployment/ Model use

- **S3: compute generality.**

$$Generality(t) = \frac{freq_g(t)}{freq_d(t)} = \frac{\frac{Occ_g(t)}{TOcc_g}}{\frac{Occ_d(t)}{TOcc_d}}$$

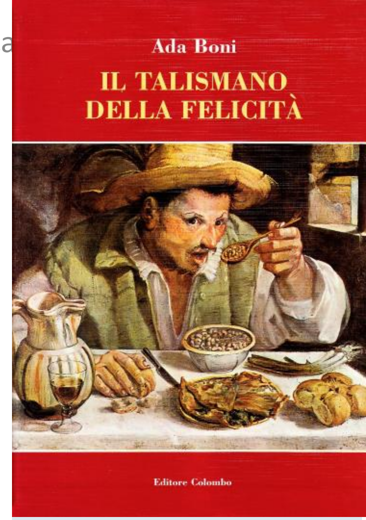
- where freq(t) is the frequency of term t, Occ(t) is the occurrence number and TOcc is the terms number-These values are computed for general domain (g) and specific domain (d).

The experiments

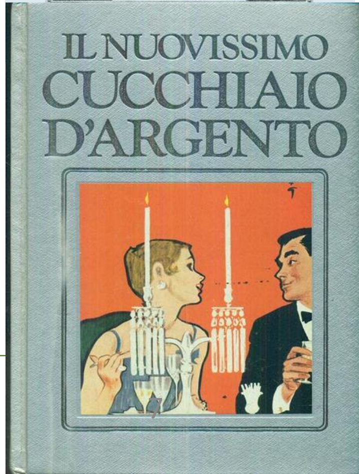
- Our primary goal is to create a list of terms related to food, in a broad vision, including ingredients, tools, actions, and tradition-related terms, in the Italian language

- Food products, diet, processing, recipes are an integral part of the cultural identity of people and communities

Intangible Cultural Heritage



PELLEGRINO ARTUSI
LA SCIENZA IN CUCINA
E
L'ARTE DI MANGIAR BENE
MANUALE PRATICO PER LE FAMIGLIE
Igiene • Economia • Buon gusto
IN FIRENZE
PEI TIPI DI SALVADORE LANDI
1891



Datasets

- CookIT, a web portal aimed at collecting, and sharing Italian traditional recipes related to regional cuisine
http://arm.mi.imati.cnr.it/cookIT

- Web pages scraped from Wikipedia, starting from the 9 root categories, plus the food-related category

- Human-created list of terms, containing ingredients, tools, and actions, taken from the web and manually integrated with missing terms.

- specific terms (1-gram and n-grams/ nouns, nouns+adjectives and verbs) from free available recipes e-books
general terms and their occurrences have been taken from the web

Preliminary results - 1

- 2 classification algorithms:
  - Logistic Regression and K-nearest neighbors
- 2 word embedding models:
  - Word2Vec and GloVe -- Pre-trained for Italian

- Several ebooks freely available:
  - Artusi
  - Cookaround, La Stampa, …
- List of most used words in Italian: about 800.000 words with their frequencies

https://github.com/hermitdave/FrequencyWords/tree/master/content/2018

Preliminary results - 2

From food related datasets:

- Knn 2 and Knn 5 extract higher number of terms with respect to Logistic Regression
- Generality helps in strengthening membership belonging to the domain.

From general datasets:

- Logistic Regression classifier performs best in this case.
- Generality helps in eliminating terms erroneously inserted in the list

Conclusions

- Work in progress
- ML classification methods on pre-trained word embedding models
- Datasets in Italian language
- Preliminary results are encouraging

- Easily adaptable to other contexts or languages

Maria Teresa Artese teresa@mi.imati.cnr.it

Isabella Gagliardi gagliardi@mi.imati.cnr.it

http://www.imati.cnr.it

http://arm.mi.imati.cnr.it/mislab/home.htm

http://arm.mi.imati.cnr.it/midb

CookIT Online Archive || recipes and images of Italian traditional Cuisine

http://arm.imati.cnr.it/cookIT

Research project developed as collaboration between IMATI – CNR with University of Milan Bicocca - DISCo Imaging and Vision Lab