!Translate : When You Cannot Cook Up a Translation, Explain

Federico Garcea[1], **Margherita Martinelli**[2],
Maja Miličević Petrović[1] and **Alberto Barrón-Cedeño**[1]
[1] Università di Bologna, Forlì, Italy
[2] Stadler, Bussnang, Switzerland
[federico.garcea2, maja.milicevic2, a.barron]@unibo.it
martinellimargherita1997@gmail.com

Abstract

In the domain of cuisine, both dishes and ingredients tend to be heavily rooted in the local context they belong to. As a result, the associated terms are often *realia* tied to specific cultures and languages. This causes difficulties for non-speakers of the local language and machine translation (MT) systems alike, as it implies a lack of the concept and/or of a plausible translation. MT typically opts for one of two alternatives: keeping the source language terms untranslated or relying on a hyperonym/near-synonym in the target language, provided one exists. !Translate proposes a better alternative: explaining. Given a cuisine entry such as a restaurant menu item, we identify culture-specific terms and enrich the output of the MT system with automatically retrieved definitions of the non-translatable terms in the target language, making the translation more actionable for the final user.

1 Introduction

National and regional cuisines are heavily tied to their historical and socio-cultural background (Civitello, 2011). Ingredients are often used differently within different cultures (e.g., whereas *hibiscus* represents a spice for chicken soup in the Philippines, it is the main ingredient for a fresh drink in Mexico[1]). Sometimes, an ingredient is widely present, but is used only in a specific region (e.g., *stridoli*[2] grow across Europe, but only some varieties are edible and are used primarily in Italian cuisine). Geographical and cultural diversity have led to the creation of unique local recipes that have no equivalents elsewhere; e.g., *strozzapreti* (an Italian pasta type) and *shish kebab* (a Middle East grilled meat dish) are not available in other cultures and, as a result, are not translated into other languages. In translation studies, such cases fall under *realia*, words referring to objects of the local material culture associated with a lack of the relevant concept and/or of a plausible translation in other languages (Vlakhov and Florin, 1970). In human translation, realia are often left untranslated (transcribed, transliterated or adapted according to the norm of the target language), and can in addition be explained by the translator, in notes or directly in the text (Florin, 1993). In MT, the problem of untranslatable items is solved either by keeping the realia untranslated, or by translating them with a hyperonym or a near-synonym in the target language.

In this demo, we focus on realia in Italian cuisine. This is one of the most widespread cuisines in the world (Capatti and Montanari, 2005), whose most dishes lack a translation in other languages, and are instead denoted by the original Italian vocabulary. Leaving aside items turned international, such as *pizza* or *cappuccino*, this phenomenon can produce a negative effect on non-Italian speakers, who might struggle to understand the meaning of most dishes and ingredients.

Our !Translate system (a) prevents a machine translation system from attempting to translate non-translatable terms, and (b) enriches the resulting partial translation with definitions of such non-translatable items, which are automatically identified and extracted from encyclopedic articles, in order to increase overall text comprehensibility.[3]

The paper is organised as follows. Section 2 introduces our approach to the identification of non-translatable fragments. Section 3 describes our method for the supervised retrieval of definitions. Section 4 outlines the architecture of the !Trans-

[1] Compare https://en.wikipedia.org/wiki/Hibiscus and https://es.wikipedia.org/wiki/Hibiscus

[2] https://it.wikipedia.org/wiki/Silene_vulgaris

[3] Prototype available at https://nt.dipintra.it

Proceedings of Recent Advances in Natural Language Processing, pages 392–398
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_044

Italian categories		
antipasti	secondi piatti	contorni
primi piatti	piatti unici	dolci

English categories	
Italian cuisine	C. of Abruzzo
C. of Apulia	C. of Basilicata
C. of Calabria	C. of Campania
C. of Emilia-Romagna	C. of Lazio
C. of Liguria	C. of Lombardy
C. of Marche	C. of Molise
C. of Piedmond	C. of Sardinia
C. of Sicily	C. of South Tyrol
C. of Tuscany	C. of Umbria
C. of Veneto	C. of Aosta Valley
Neapolitan cuisine	Italian desserts

Table 1: Wikipedia categories considered as relevant for the Italian cuisine in both the Italian and English (C.=Cuisine).

late system. Section 5 overviews related work. Section 6 closes with conclusions and further work.

2 Identification of Non-Translatable Fragments

Sentences that contain terms or phrases that are out of vocabulary for an MT engine typically yield low-quality MT output. Hence, we can use a list of entries (glossary) for regional dish names and ingredients, and adopt a brute force approach to identify non-translatable fragments. We iterate through the glossary in the source language and find the longest match in the input sentence. By using the longest match, we take advantage of glossary entries that may contain the full name of a traditional dish, as opposed to single words for a specific ingredient.

The matching algorithm considers variants of a term, i.e. aliases that are contained in each glossary entry, since it is common for regional dishes to have more than one name (usually because the original name was in a local dialect and has since been 'italianised', taking a slightly different form), and either variant can appear in restaurant menus or recipes. While more sophisticated entity-linking models could be used (cf. Section 5), this brute-force approach proved to be enough in the cuisine setting.

Our glossary is built from Wikipedia entries that belong to categories associated with the Italian cuisine and from an in-house parallel collection of regional-cuisine menu entries prepared by professional translators.[4] To select the subset of relevant Wikipedia articles both in Italian and English, we rely on the categorisation of the Wikipedia itself and select those entries that belong to, at least, one of the relevant categories. Table 1 shows the categories used for the two languages. As expected, there are very few parallel categories for the cuisine domain (*dolci* and *Italian desserts*), which reflects the standpoint of the Wikipedia editions in the two languages.

	P	**R**	**F_1**
Wikifier	23.44	**54.05**	32.70
Brute force	**88.06**	53.15	**66.29**

Table 2: Performance of the alternatives for the identification of non-translatable fragments.

In order to assess the performance of the alternative approaches to non-translatable fragments identification, three annotators labelled 120 instances —one native speaker of Italian and two advanced non-native speakers. After consolidation, 111 text spans were identified as non-translatable. Table 2 shows the performance of two alternative models: our brute-force approach and a standard entity linking approach (Brank et al., 2017). Whereas the recall values are comparable for both models, the precision of our approach is more than three times better, boosting the F_1-measure. This is thanks to the applied glossary, which prevents the model from greedily identifying all (pseudo-)terms.

3 Acquisition of Definitions

In order to obtain the necessary definitions, we aim at automatically extracting definitional contexts from the Wikipedia, the largest multilingual collection of copyright-free encyclopedic content. We use the Italian and English Wikipedia dumps from July 2021 and keep only the articles that belong to the Italian cuisine, according to their associated categories (cf. Table 1 for the whole list of categories). Table 3 shows statistics of the resulting dataset, which displays the expected distribution: more articles in Italian about the Italian cuisine, even if the articles tend to be longer in English.

Our objective is extracting definitional contexts that can explain non-translatable cuisine terms

[4] Professional translations from Italian into English of the menus from the 2021 edition of the *Festa Artusiana*, a regional cuisine festival (http://www.festartusiana.it).

	it	en
articles	2,054	1,923
tokens	780,996	1,170,360
avg. length	380	608

Table 3: Statistics of the articles associated to the Italian cuisine identified in the Italian and English editions of the Wikipedia (avg. article length computed in tokens).

Gnudi are gnocchi-like dumplings made with ricotta cheese instead of potato, with semolina.
The result is often a lighter, “pillowy” dish, unlike the often denser, chewier gnocchi.
Gnudi is the Tuscan word for ”naked” (in standard Italian “nudi”), the idea being that these “pillowy” balls of ricotta and spinach (sometimes without spinach, which is also known as ricotta gnocchi) are “nude ravioli”, consisting of just the tasty filling without the pasta shell.
By tradition, in Tuscany, these dumplings are served with burnt butter and sage sauce, sprinkled with Parmigiano or Pecorino Toscano cheese.
. . .

Figure 1: A Wikipedia article (input) with its definitional context framed (output), as identified by the BERT-based model.

across languages. Aristotle formulated definitional contexts as sequences of type

$$X = Y + C \ , \qquad (1)$$

where X is the *definiendum* (the term), $=$ is the *definitor* (a connective verb such as ‘to be’ or ‘consist’), Y is the *definiens* (the genus phrase, or nearest superconcept), and C are the *differentiæ specificæ*, the distinguishing characteristics that specify the distinction between one definiendum and another (Del Gaudio et al., 2014). For example, the definitional context for *gnudi* is as follows:

$$\overbrace{\text{Gnudi}}^{X} \ \overbrace{\textit{are}}^{=} \ \overbrace{\text{gnocchi-like dumplings}}^{Y}$$
$$\underbrace{\text{made with ricotta cheese instead of potato}}_{C}$$

In order to train the model to identify such definitional contexts, we use the corpus produced by Navigli et al. (2010). It is a collection of 4,719 items, each containing the opening sentences of a Wikipedia article in English. Definitional contexts in this collection were manually identified, resulting in 1,872 positive instances. Figure 1 shows an example of the input —a full Wikipedia article— with the expected output.

	F_1	Acc
Navigli and Velardi (2010)*	75.23	83.84
bert-base-cased	96.08	96.82
bert-base-multilingual-cased	**97.66**	**98.09**

*No official testing partition has been published; hence these numbers are not directly comparable against ours.

Table 4: Performance of the two model variations for the identification of definitional contexts.

We experimented with two models based on BERT (Devlin et al., 2019) to classify sentences as definitional context or not: `bert-base-cased` and `bert-base-multilingual-cased`. The former is intended for the extraction of definitions when the target language is English, whereas the latter is intended to give an estimation of the performance when requiring definitions in Italian. We split the dataset into 80% for training, 10% for validation and 10% for testing. Table 4 shows the performance obtained on the testing partition. The performance of both the monolingual and the multilingual alternatives is remarkable, landing close to a perfect accuracy.

Table 5 shows some examples of definitional-context candidates that our model identifies in Wikipedia articles, both in English and Italian. Both instances 1 and 3 represent proper definitional contexts that would help a user to understand a dish. Instance 2 is a proper definitional context, but with a clear encyclopedic spirit. Instance 4 refers to the story of fish fingers rather than a proper definition.

4 The !Translate Components

Figure 2 illustrates the architecture of the !Translate system, which is composed of the backend and the frontend.

Backend The *backend* website allows project contributors to manage glossaries and their entries. The *multilingual glossary* itself is a database that is accessed through APIs by the backend website and the *definition extractor* component. Not all cuisine-related entries in the Italian Wikipedia have a corresponding page in English. For those, we use MT to translate the best definition extracted from the Italian page.

Frontend The *frontend user interface* is a website that accepts user input and displays enhanced translations in the desired language. The input is a free text (e.g., a recipe, a restaurant menu) which is

definitional context	op
English	
1. **Picada** is a type of tapas eaten in Argentina and Uruguay, usually involving only cold dishes, such as olives, ham, salami, mortadella, bologna, different types of cheese, marinated eggplants and red pimentos, sardines, nuts, corn puffs, fried wheat flour sticks, potato chips, and sliced baguette	♣
2. **Sucrose** is a disaccharide made up of glucose and fructose.	
Italian	
3. I **canéderli** (in tedesco Semmelknödel) sono degli Knödel (grossi gnocchi) composti di un impasto a composizione variabile di pane raffermo.[a]	♣
4. Le **“dita di pesce”** (fish fingers) furono una ricetta di inizio Novecento pubblicata su una popolare rivista britannica ed è tuttora considerato spesso un piatto-simbolo della cucina del Regno Unito.[b]	

[a] `Canérdeli (in German Semmelknödel) are Knödel (large gnocchi) made of a dough with diverse mixtures of sourdough bread.`

[b] `Fish fingers were a recipe from the early 20th century published on a popular British magazine and is still often considered a signature dish of UK cuisine.`

Table 5: Examples of extracted definitional contexts in English (top) and Italian (bottom; English translations included for comprehensibility). Column **op** flags definitions considered operational for the !Translate explanation purposes.

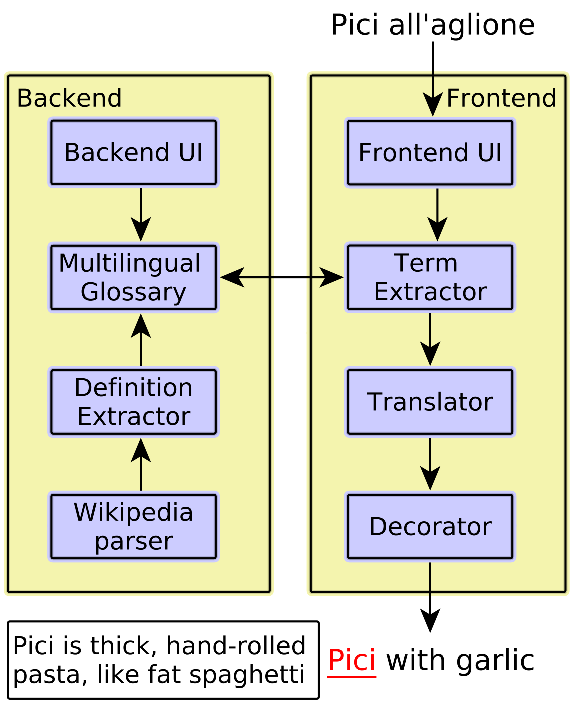


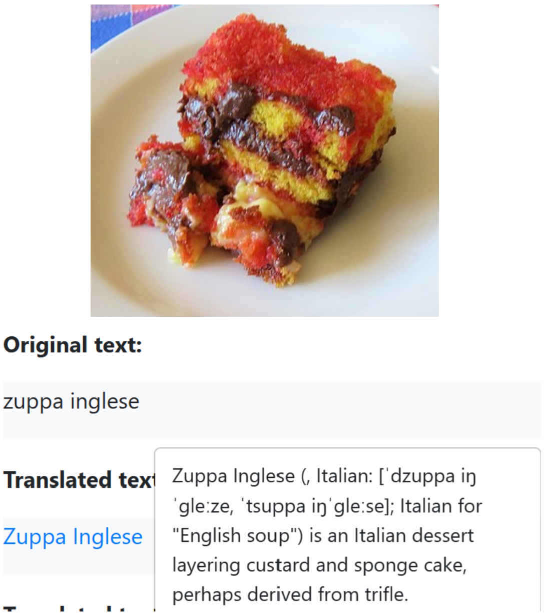
Figure 2: The !Translate system architecture

passed through a *segmentation* or sentence-breaker component to divide input text into individual sentences. A *term extractor* matches non-translatable fragments against the *multilingual glossary* and replaces them with special do-not-translate XML tags, with attributes to encapsulate the desired substitution terms. This step produces an out-of-vocabulary, preventing an MT system from attempting to translate literally certain terms and contains metadata to inform further components in the pipeline about the non-translatable items found. The *translator* component handles calls to a cloud MT engine, such as ModernMT;[5] this is a simple proxy for an online MT, with no customization or adaptation. The post-processing *decorator* component takes the MT output and, by looking at the metadata in each do-not-translate tag, substitutes these tags with a hyperlink to a definition, and their content (the fragment within do-not-translate tags) with the proper translation taken from the glossary.

As observed in the example of Figure 2, given the input *Pici all’aglione*, the system matches **Pici** with a non-translatable entry from the glossary, retrieves the pre-obtained definition, and plugs it in next it in the enhanced translated output.

Figure 3 shows a snapshot from our system.

[5] `https://github.com/modernmt/modernmt`



Original text:

zuppa inglese

Translated tex

Zuppa Inglese

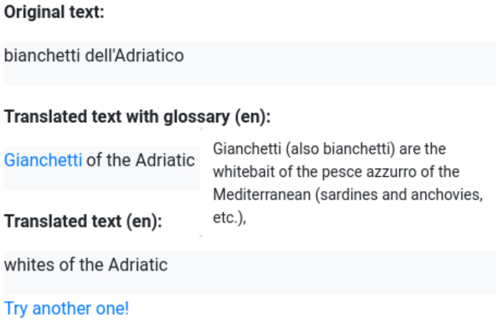
Zuppa Inglese (, Italian: [ˈdzuppa iŋ ˈɡleːze, ˈtsuppa iŋˈɡleːse]; Italian for "English soup") is an Italian dessert layering custard and sponge cake, perhaps derived from trifle.

Figure 3: A snapshopt of the system interface showing *zuppa inglese* —which is not English and is not a soup— and its augmented (no) translation.

Rather than translating the entry and providing a useless "accurate" translation ('English soup'), our system opts for keeping the entry untranslated and providing a definition instead, which properly describes the concept. Figure 4 shows another example. This time, part of the item is translated whereas another part is not, and it is explained instead: *bianchetti* are not *little whites*, but young blue fish, such as sardines.

5 Related Work

Entity linking aims at identifying the unique identity of an entry. This kind of technology is commonly supported on linking text to encyclopedic entries. Such a process is also known as wikification, in which the entities are linked to the Wikipedia in order to augment the comprehensibility of a text. One of the first approaches was Wikify! (Mihalcea and Csomai, 2007), which relied on a combination of steps to perform keyword-matching and disambiguation independently. Babelfy (Moro et al., 2014) is another alternative, but its approach to word disambiguation targets to identify all concepts which, for our purposes, results in over-identification. In recent approaches, entity linking is modeled with neural models that perform the task of entity finding and linking at once (Kolitsas et al., 2018). Through a dual encoder, the model proposed by Botha et al. (2020) can link entities in multiple languages. We do not opt for any of these models because the texts we deals with are brief (e.g., menu entries) and rather than performing an open search, we only need to find matches.



Original text:

bianchetti dell'Adriatico

Translated text with glossary (en):

Gianchetti of the Adriatic

Gianchetti (also bianchetti) are the whitebait of the pesce azzurro of the Mediterranean (sardines and anchovies, etc.),

Translated text (en):

whites of the Adriatic

Try another one!

Figure 4: A snapshopt of the system interface showing *bianchetti dell'Adriatico*. At the bottom the default (wrong translation). In the middle, the correct and augmented partial translation: 'Gianchetti of the Adriatic'.

The task of extracting definitional contexts is not limited to glossaries and encyclopaediae, but extended to other fields such as ontology learning (Gangemi et al., 2003), question answering (Saggion, 2004; Cui et al., 2007) and eLearning (Westerhout and Monachesi, 2007). Most approaches rely on lexico–syntactic patterns (Saggion, 2004; Cui et al., 2007; Fahmi and Bouma, 2006; Degórski et al., 2008) that require manual annotation and/or manually written rules. A different approach has been taken with the use of Word Lattices, directed acyclic graphs that represent a segment. (Navigli and Velardi, 2010) introduced Word-Class Lattices to model textual definitions.

6 Conclusions and Future Work

We have presented !Translate , an application that automatically produces translations combining machine translation, entity linking, and supervised definition retrieval to provide informative translations to users in settings in which machine translation alone is not enough. We have focused on the domain of cuisine, in which terms often lack in the target language and require further descriptions (definitions) to become operational.

As part of our ongoing work, we are experimenting with a MT Quality Estimation (QE) component to optionally direct the translation request to a notification queue component that will post a request to a crowdsourcing-based translation component for those sentences that are deemed difficult to translate automatically, even with the help of a glossary.

Ethics/Broader Impact

This paper presents a system that enhances machine translation via automatic identification of untranslatable terms and automatic extraction of definitions for these terms, which are then added to the MT output. Our focus is on culture-specific items in restaurant menus written in Italian, but our pipeline may benefit applications dealing with other specialised domains. On a wider societal plan, our work concerns intangible cultural heritage and aims to help protect local traditions by using local names while at the same time explaining their meaning to those who might not be familiar with them. We do not see any potential for malicious usage of our framework.

Acknowledgments

This research work was carried out within the framework of Università di Bologna's Alma Idea 2022 project !Trans, financed by *Decreto Ministeriale* 737/2021 of the Italian Government.

References

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Ljubljana, Slovenia.

Alberto Capatti and Massimo Montanari. 2005. *La cucina italiana. Storia di una cultura*. Laterza, Bari.

Linda Civitello. 2011. *Cuisine and Culture: A History of Food and People*. Wiley, Hoboken, New Jersey.

Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2007. Soft pattern matching models for definitional question answering. *ACM Trans. Inf. Syst.*, 25(2):8–es.

Łukasz Degórski, Michał Marcińczuk, and Adam Przepiórkowski. 2008. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rosa Del Gaudio, Gustavo Batista, and Antonio Branco. 2014. Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, 20(3):327–359.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ismail Fahmi and Gosse Bouma. 2006. Learning to identify definitions using syntactic features. In *Proceedings of the Workshop on Learning Structured Information in Natural Language Applications*.

Sider Florin. 1993. Realia in translation. In Palma Zlateva, editor, *Translation as Social Action: Russian and Bulgarian Perspectives*, pages 122–128. Routledge, London and New York.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pages 820–838, Berlin, Heidelberg. Springer Berlin Heidelberg.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233—242, New York, NY, USA. Association for Computing Machinery.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327, Uppsala, Sweden. Association for Computational Linguistics.

Roberto Navigli, Paola Velardi, and Juana Maria Ruiz-Martínez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Horacio Saggion. 2004. Identifying definitions in text collections for question answering. In *Proceedings of* *the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sergei Vlakhov and Sider Florin. 1970. Neperevodimoye v perevode: realii. *Masterstvo perevoda*, 6:432–456.

Eline Westerhout and Paola Monachesi. 2007. Extraction of Dutch definitory contexts for eLearning purposes. *LOT Occasional Series*, 7:219–234.