English Version (Italian Below)

!Trans: when translating is not enough, explain

Introduction

Project !Trans (no-translate) is developed within the framework of Alma Idea 2022, an initiative of Alma Mater Studiorum-Università di Bologna in agreement with the Italian PNR (National Recovery Plan). !Trans aims at developing technology to spot texts for which even modern machine translation technologies run short due to terms/domains heavily rooted in a region/culture, causing valid translations to hardly exist in other languages. !Trans aims at developing technology to explain concepts across languages in order to increase comprehensibility under such an unfavourable scenario.

Objectives and Expected Results

Main objective. Developing (semi-)automatic pipelines to explain complex or untranslatable terms, specific to the intangible Italian gastronomy heritage, to enhance the comprehension and evaluate the usability of texts in (incidental) learning contexts for Italian as L2.

Specific objectives.

1. Design and develop supervised models to assess the level of (cross-language) comprehensibility of a text, as well as the feasibility of its automatic translation [1].
2. Design and develop supervised models for the retrieval of cross-language definitions to produce explanations of complex/specific/non-translatable terms from a culture/language. [2]

The project is devoted to the preservation of one of the least known intangible human heritage: gastronomy. The target users are different kinds of visitors (tourists, exchange students), but it can also impact other user types, as far as they are non-native speakers of the source language --Italian.

The practical problem concerns various scenarios in which a non-native speaker (e.g., a tourist or an exchange student) faces difficulties to understand a text due to the limited command of a specialised vocabulary (e.g., the terminology of gastronomy). Often, relying on (machine) translation is not optimal since, for a number of terms, no translation exists at all. In such a scenario, an alternative solution becomes necessary: displaying the definition/description of the term (e.g., a menu entry, such as *strozzapreti* or *quinto quarto*), in Italian or another language (English or the user's native language).

Outline

1. Producing and organising in-domain multilingual corpora

Acquisition of comparable in-domain texts in both Italian and English (e.g., gastronomic tradition, international student guides). The datasets will be organised with the help of one or more automatic models for the extraction and classification of the texts.

The expected result is a model for the classification of in-domain multilingual corpora.

2. Developing supervised models to assess the (inter-)comprehensibility of a text and the feasibility of producing a correct translation out of it.

In this phase we will develop a system to assess the complexity of a text and its level of *translatability*, understood as an estimation of the quality of the machine translated version in the target language (e.g., from Italian to English).

The best model will be used in a hybrid translation prototype where texts will be sent to a machine translation system or to a manager of human translation requests.

The expected outcome is a software prototype for the management of translation requests, including a quality estimation (QE) module.

3. Development of supervised models for the retrieval of multilingual definitions in order to explain complex/specific/untranslatable terms from the source culture/language.

Within this third and final phase, we will experiment with models to spot terms that should be defined rather than translated.
It will be followed by the research and development of methods for the automatic extraction of definitions from ad hoc dorpora produced from reference in-domain sources.
Finally, we will produce an interface prototype to acquire texts and return an “enriched” version, including the definitions for the automatically spotted untranslatable terms.

The expected outcome is a prototype gathering together all the models and texts produced during the different project phases, which are going to represent the ideal experience for the acquisition, transformation, and visualisation of texts with untranslatable terms.

Incoming Profile

Master Degree / specialist or old system in computer science or specialised translation and interpreting, with appropriate scientific and professional curriculum.
The ideal candidate has a strong background in natural language processing / computational linguistics, in particular with regards to machine learning and machine translation. The candidate must possess knowledge of machine learning models (deep learning) applied to natural language processing, and the ability to implement them through standard libraries (e.g., scikit, TF, Keras, spacy, huggingface).
They also have the ability to create customised machine translation engines, and to understand and apply methods for quality evaluation and quality estimation of MT output.
Prior experience in creating prototypes of mobile or web applications is required, as well as measuring and understanding user feedback.
The candidate must also have near native knowledge of Italian and English.

Training and Supervision

The successful candidate will work within an interdisciplinary group with backgrounds in natural language processing and human translation. Alberto Barrón-Cedeño (https://www.unibo.it/sitoweb/a.barron) is the PI. He is a Computing Scientist with expertise in (multilingual) text analysis. Maja Milicevic Petrovic (https://www.unibo.it/sitoweb/maja.milicevic2) is the co-PI. She works on linguistic aspects of different bilingual situations and different linguistic varieties. The successful candidate will follow a learning-by-doing approach in which (s)he will develop resources and models under the close collaboration with both supervisors, paying attention to both the computational and linguistic side of the problem.

References

Fernicola, F. (2022). Return to the Source: Assessing Machine Translation Suitability based on the Source Text using XLM-RoBERTa. Tesi di Laurea Magistrale. Dipartimento di Interpretazione e Traduzione. Alma Mater Studiorum-Università di Bologna.

Forti, L., Milani, A., Piersanti, L., Santarelli, F., Santucci, V., and Spina, S. (2019). Measuring text complexity for Italian as a second language learning purposes. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 360-368, Florence, Italy, August. Association for Computational Linguistics.

Martinelli M. (2022) Methods of Definition Extraction and Linking for Food Recipes. Tesi di Laurea Magistrale. Dipartimento di Interpretazione e Traduzione. Alma Mater Studiorum-Università di Bologna.

!Trans: quando la traduzione non è sufficiente, spiega[1]

Introduzione

Il progetto !Trans (no-translate) è sviluppato nell'ambito di Alma Idea 2022, un'iniziativa dell'Alma Mater Studiorum-Università di Bologna in accordo con il PNR italiano. !Trans vuole a sviluppare tecnologia per individuare i testi per i quali anche le moderne tecnologie di traduzione automatica sono insufficienti a causa di termini/domini fortemente radicati in una regione/cultura, per cui difficilmente esistono traduzioni validi in altre lingue. !Trans mira a svilupperà tecnologia per spiegare i concetti tra le varie lingue, al fine di aumentare la comprensibilità in uno scenario così sfavorevole.

Obiettivi e risultati attesi

Obiettivo principale. Sviluppare delle pipeline automatiche per la spiegazione translingue di termini complessi e/o intraducibili, specifici del patrimonio intangibile italiano nell'ambito della gastronomia, allo scopo di migliorare la comprensione dei testi e valutarne l'usabilità in contesti di apprendimento (anche incidentale) dell'italiano L2.

Obiettivi specifici.

1. Progettare e sviluppare modelli supervisionati per la valutazione della (inter-)comprensibilità di un testo e la fattibilità della sua traduzione automatica. [1]
2. Progettare e sviluppare modelli supervisionati per il recupero automatico di definizioni multilingue per le spiegazioni di termini complessi/specifici/intraducibili della cultura/linguaggio fonte, valutando automaticamente la complessità testuale della definizione e la sua adeguatezza nell'ottica di italiano L2. [2]

Il progetto è dedicato alla conservazione, tramite l'utilizzo di tecnologie, del patrimonio culturale intangibile meno conosciuto, quale la tradizione culinaria. I destinatari principali sono diverse tipologie di visitatori (turisti, studenti in scambio), ma anche altre tipologie di utenti interessati agli ambiti coperti dal progetto. La caratteristica condivisa è lo status di parlante non nativo di lingua italiana.

Il problema pratico che si vuole risolvere riguarda i diversi scenari in cui un parlante non nativo (per. esempio, un turista o una studentessa in scambio) ha problemi di comprensione di un testo per via delle lacune nella conoscenza del vocabolario specializzato (per es. nella terminologia legata alla gastronomia). In molti casi, non sarà sufficiente appoggiarsi alla traduzione (automatica), visto che per alcuni termini una traduzione non esiste. In questo caso, bisogna trovare una soluzione diversa, che può essere una definizione/descrizione del

[1] ! all'inizio di Trans(lation) ha il ruolo di una negazione: "no-translation".

termine problematico (per esempio, un elemento di un menù, come *strozzapreti* o *quinto quarto*), in italiano o in una lingua diversa (l'inglese o la lingua madre della persona interessata).

Piano dell'attività

1. Creazione e organizzazione di corpora multilingue in-domain

In questa fase, verrà fatta una acquisizione di dati testuali comparabili in italiano e in inglese, con la ricerca e selezione di testi per i domini linguistici che ci interessa esplorare (es. tradizione culinaria, guide per lo studente internazionale).
I dati verranno organizzati anche con l'aiuto di uno o più modelli automatici per l'estrazione e la classificazione del testo.

Il risultato atteso è un modello di classificazione e corpora multilingue di dati per ogni dominio esplorato.

2. Sviluppo di modelli supervisionati per la valutazione della (inter-)comprensibilità di un testo e la fattibilità della sua traduzione automatica.

In questa fase svilupperemo un sistema in grado di valutare la complessità di un testo e la sua *traducibilità*, intesa come stima della qualità della traduzione automatica in un'altra lingua di riferimento (es. da italiano a inglese).

Successivamente, il modello migliore viene utilizzato in un prototipo di sistema ibrido di traduzione in cui i testi possono essere inviati a un sistema automatico di traduzione o a un sistema di gestione della traduzione umana.

Il risultato atteso è un prototipo di software della gestione delle traduzioni, con un modulo di stima della qualità prevista (QE).

3. Sviluppo di modelli supervisionati per il recupero automatico di definizioni multilingue per le spiegazioni di termini complessi/specifici/intraducibili della cultura/linguaggio fonte.

In questa terza ed ultima fase, intendiamo sperimentare con modelli per l'identificazione di termini da definire e non tradurre.
Seguirà la ricerca e sviluppo di metodi per l'estrazione automatica di definizioni da corpora appropriatamente creati ed estratti da fonti rilevanti nei domini di interesse.
Infine, bisognerà progettare e realizzare un prototipo di interfaccia per acquisire un testo e mostrarne la versione 'arricchita' con le definizioni dei termini 'intraducibili' selezionati automaticamente.

Il risultato atteso è un prototipo di una applicazione che metta insieme i modelli e testi prodotti da tutte le fasi del progetto, e rappresenti l'esperienza ideale per l'acquisizione, trasformazione, e visualizzazione di testi contenenti termini intraducibili.

Profilo in entrata

Laurea magistrale/specialistica o vecchio ordinamento in informatica oppure traduzione specialistica e interpretariato con adeguato curriculum scientifico-professionale.
Il candidato ideale ha competenze avanzate in natural language processing / computational linguistics, in particolare nell'ambito di machine learning e machine translation.
Il candidato possiede conoscenze dei modelli di machine learning (deep learning) applicati al natural language processing, e la capacità di implementarli attraverso librerie standard (ad es. scikit, TF, Keras, spacy, huggingface).
Il candidato ha la capacità di creare motori di traduzione automatica personalizzati, e di comprendere ed applicare metodi di valutazione della qualità e stima della qualità dell'output dei motori di traduzione.
Inoltre, è necessario avere esperienza nella realizzazione di prototipi di applicazioni web e mobili, e metodi di misurazione della soddisfazione degli utenti.
Si richiede infine conoscenza della lingua italiana se cittadino straniero e della lingua inglese a livello C1.

Formazione/Supervisione

Il candidato vincitore lavorerà all'interno di un gruppo interdisciplinare con background nell'elaborazione del linguaggio naturale e la traduzione. Alberto Barrón-Cedeño (https://www.unibo.it/sitoweb/a.barron) è il PI. È uno scienziato informatico con esperienza nell'analisi del testo (multilingue). Maja Milicevic Petrovic (https://www.unibo.it/sitoweb/maja.milicevic2) è la co-PI. Lavora sugli aspetti linguistici delle diverse situazioni bilingui e delle diverse varietà linguistiche.

Il candidato vincitore seguirà un approccio "learning-by-doing" in cui svilupperà risorse e modelli in stretta collaborazione con entrambi i supervisori, prestando attenzione sia all'aspetto computazionale che linguistico del problema.

Bibliografia

Fernicola, F. (2022). Return to the Source: Assessing Machine Translation Suitability based on the Source Text using XLM-RoBERTa. Tesi di Laurea Magistrale. Dipartimento di Interpretazione e Traduzione. Alma Mater Studiorum-Università di Bologna.

Forti, L., Milani, A., Piersanti, L., Santarelli, F., Santucci, V., and Spina, S. (2019). Measuring text complexity for Italian as a second language learning purposes. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 360-368, Florence, Italy, August. Association for Computational Linguistics.

Martinelli M. (2022) Methods of Definition Extraction and Linking for Food Recipes. Tesi di Laurea Magistrale. Dipartimento di Interpretazione e Traduzione. Alma Mater Studiorum-Università di Bologna.