Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1–12
Marseille, 20-25 June 2022

Domain Adaptation in Neural Machine Translation using a Qualia-Enriched FrameNet

Alexandre Diniz Costa, Mateus Coutinho Marim, Ely Edison da Silva Matos, Tiago Timponi Torrent

Federal University of Juiz de Fora | FrameNet Brasil
Rua José Lourenço Kelmer, s/nº – Campus Universitário – Juiz de Fora, MG – Brazil
{alexandre.costa,ely.matos,tiago.torrent}@ufjf.br, mateus.marim@ice.ufjf.br

Abstract

In this paper we present Scylla, a methodology for domain adaptation of Neural Machine Translation (NMT) systems that make use of a multilingual FrameNet enriched with qualia relations as an external knowledge base. Domain adaptation techniques used in NMT usually require fine-tuning and in-domain training data, which may pose difficulties for those working with lesser-resourced languages and may also lead to performance decay of the NMT system for out-of-domain sentences. Scylla does not require fine-tuning of the NMT model, avoiding the risk of model over-fitting and consequent decrease in performance for out-of-domain translations. Two versions of Scylla are presented: one using the source sentence as input, and another one using the target sentence. We evaluate Scylla in comparison to a state-of-the-art commercial NMT system in an experiment in which 50 sentences from the Sports domain are translated from Brazilian Portuguese to English. The two versions of Scylla significantly outperform the baseline commercial system in HTER.

Keywords: FrameNet, Qualia Relations, Sports, Machine Translation, Domain Adaptation

1. Introduction

Neural models have been advancing the state-of-art in the field of Machine Translation (MT) in a myriad of tasks (Barrault et al., 2019; Barrault et al., 2020). Nonetheless, as pointed out by Koehn and Knowles (2017), maintaining the high performance of Neural Machine Translation (NMT) in a specific domain for which there is a lack of large training data is challenging. Domain adaptation has been used as a strategy to mitigate such a loss in performance.

Chu and Wang (2018) pointed out that a big research question still to be answered is how to use external knowledge such as dictionaries and knowledge bases for domain adaptation in NMT. In this paper, we provide an answer to such a question in the form of Scylla, a methodology for domain adaptation of NMT systems. Scylla substitutes domain-specific terms in the source language by their adequate translation in the target language and is implemented in two versions: one of them, Scylla-S, does it before the source sentence is fed into the NMT system, while the other, Scylla-T, takes as input the already translated sentence.

Both pipelines make use of a multilingual FrameNet covering the Sports domain (Costa and Torrent, 2017; Costa et al., 2018) as an external knowledge base. This is to say that they do not require any fine-tuning of the NMT system, avoiding model over-fitting and decrease in performance for general-domain translations (Khayrallah et al., 2018; Thompson et al., 2019).

To evaluate Scylla, we conducted an experiment where 50 Brazilian Portuguese (br-pt) sentences, collected from Sports news and encyclopedias, were submitted to the commercial NMT system alone - considered as the baseline -, and to the two pipelines presented in this paper for translation into English (en). Systems' performances were then evaluated against BLEU (Papineni et al., 2002), TER and HTER (Snover et al., 2006). The two domain adaptation solutions presented in this paper significantly outperform the baseline for HTER.

The contributions of this paper are two-fold:

1. We present two solutions for using a semantically structured external resource – FrameNet – for domain adaptation in NMT, both of which outperform the baseline;

2. The solutions proposed do not require fine-tuning of the NMT model, substantially reducing computational costs.

In the remainder of this paper, we survey, in section 2, recent research work focusing on the use of external resources for domain adaptation in MT. Section 3 presents the FrameNet model used in the methodology. Scylla is presented in section 4 and its evaluation is discussed in section 5.

2. Related Work

In a survey paper on domain adaptation in NMT, Chu and Wang (2018) list three works using external knowledge for domain adaptation in MT: Arthur et al. (2016), Zhang and Zong (2016) and Arcan and Buitelaar (2017). In this section we provide a summary of each of them and also of Moussallem et al. (2019) and Dougal and Lonsdale (2020), which were published after Chu and Wang (2018).

Arthur et al. (2016) use discrete translation lexicons to improve the performance of NMT systems for low-frequency words. Their solution involves both automatically learned lexicons, similar to those used for Statistical Machine Translation (SMT), and bilingual dictionaries, as well as a combination of both. Although their solution improves the performance of the NMT system for the English-Japanese language pair, it is not actually focused on domain adaptation, but on addressing the issue of low-frequency words, usually classified as unknown in NMT systems. As we will demonstrate in section 4, Scylla performs domain adaptation substitutions even for frequent lexical units known by the NMT model.

Zhang and Zong (2016) use bilingual dictionaries to generate pseudo sentence pairs that are fed into the NMT system during training. Scylla, on the other hand, does not require any dataset to be generated from the qualia-enriched FrameNet lexicon used, nor any additional training of the NMT system.

Arcan and Buitelaar (2017) compare the performance of SMT and NMT systems in the task of translating domain-specific terms out of any context. In the experiments they conducted with NMT models, they used a bilingual lexicon in the form of correspondence tables for substituting unknown words in OpenNMT (Klein et al., 2017). They report improvement in performance for some experimental setups, but, once again, the solution is focused exclusively on unknown words.

Moussallem et al. (2019) propose a methodology to incorporate Knowledge Graphs (KGs) into NMT models in two steps. First, they connect named entities in parallel corpora used for training the NMT system to a reference KG using a multilingual entity linking system. Then, they concatenate the KG embeddings into the embeddings of the NMT system. Authors report improved performance for BLEU, METEOR and CHRF3. Nonetheless, their solution requires training corpora to be annotated.

Dougal and Lonsdale (2020) present a solution where a bilingual termbase is used for substituting one word in the translated sentence by another word or expression listed as an equivalent in the termbase. The algorithm is to some extent similar to that of the Scylla-T pipeline, in which regards the identification of the substitution points in the sentence. However, the solution in Dougal and Lonsdale (2020) does not rely on a semantically-structured database capable of telling the contexts where the terminology injection should take place apart from those where it should not. Moreover, their solution cannot perform either many-to-1 or many-to-many substitutions.

3. A Qualia-Enriched FrameNet

FrameNet (Baker et al., 1998; Fillmore and Baker, 2009) is an implementation of the theory of Frame Semantics (Fillmore, 1982) in which the lexicon of the English language is modeled against a network of background scenes – or frames. Each frame is composed of frame elements (FEs), which indicate the participants and props in the scene. From the early 2000's on, the framenet model has been expanded into other languages (Baker and Lorenzi, 2020), FrameNet Brasil (FN-Br) being the Brazilian Portuguese branch of this initiative (Torrent et al., 2018).

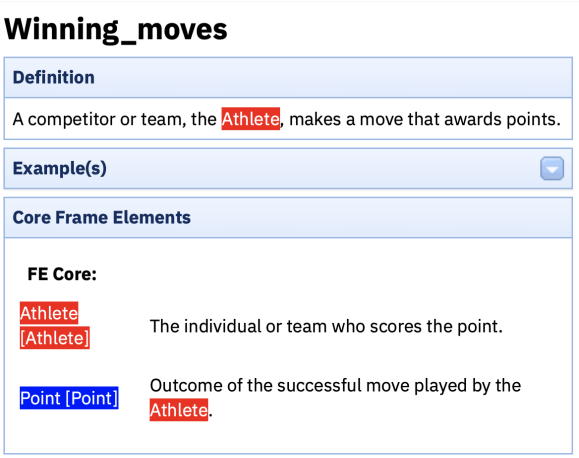

Figure 1: The `Winning_moves` frame in FN-Br

On top of expanding the framenet model into br-pt, FN-Br also develops multilingual frames for specific domains, such as Tourism and Olympic Sports (Torrent et al., 2014; Costa and Torrent, 2017; Costa et al., 2018). Because the human experience in those domains tends to abide by highly internationalized standards, frames in them also tend to be cross-linguistically applicable (Torrent et al., 2014). Therefore, one same structure, such as the `Winning_moves` frame depicted in Figure 1, can be evoked by both br-pt lexical units (LUs) like *bandeja.n, gol.n* and *marcar.v* and their en equivalents *lay-up.n, goal.n* and *score.v*, respectively.

In any framenet, frames are connected to each other via typed relations such as inheritance, subframe, perspective on, using, among others (Ruppenhofer et al., 2016). The `Winning_moves` frame, for instance, inherits the `Moves` frame, since it models specific kinds of moves that result in scoring a point. In turn, `Moves` uses `Athletes`, since any LU indicating a move will to some extent make reference to the athlete performing it.

Although Frame-to-Frame relations capture important aspects of meaning in a framenet, they are not capable of representing all the semantic relations need to properly model a domain. For instance, the original FrameNet model has no means of representing that the ATHLETE FE in the `Winning_moves` frame can be defined in terms of the `Athletes` frame. This is to say that LUs evoking the latter, will most likely be the prototypical fillers of the ATHLETE FE in the former. FN-Br captures this information via a FE-to-Frame relation.

Moreover, relations that are specific for a set of LUs

	Const.	Agent.	Telic	Formal
br-pt	950	82	1,868	1,432
en	1,290	118	2,462	3,012

Table 1: TQRs created for the Sports domain

within the frames are not captured by Frame-to-Frame relations either. In the example being discussed, the relations connecting `Winning_moves` to `Athletes` via the `Moves` frame are not able to represent that a *lay up* is a winning move performed by a *basketball player*, but not by a *soccer player*. To address this issue, FN-Br uses qualia relations (Pustejovsky, 1995) to connect LUs in different frames.

The four original qualia proposed by Pustejovsky (1995) – agentive, constitutive, formal, and telic – are very general and a collection of efforts have been made to specify them (Lenci et al., 2000; Pustejovsky et al., 2006). Instead of building or relying on an external ontology for refining the meaning of qualia relations, FN-Br specifies each type of quale by resorting to a frame that mediates the quale connecting two LUs. For example, according to Pustejovsky (1995), the relation held between *lay up.n* in the `Winning_moves` frame and *basketball player.n* in the `Athletes` frame would be a telic one, since scoring the point is the intention of the athlete performing the move. In the FN-Br database, those two LUs are connected via a ternary telic relation mediated by the `Intentionally_act` frame, which features two core FEs: the AGENT and the ACTION they perform. In the ternary quale being discussed, the LU *basketball player.n* is connected to the first FE, while *lay up.n* is framed by the second[1].

The qualia-enriched framenet model of the Sports domain used by Scylla was developed by Costa (2020), and features 36 frames, which can be evoked by 1,651 LUs in br-pt and 2,051 in en. A total of 4,332 instances of TQRs were created for br-pt, being then automatically replicated for en. Because some br-pt LUs may have more than one translation equivalent in en, the replication procedure yielded a total of 6,882 instances of TQRs. The distribution of TQRs per quale and per language is shown in Table 1.

4. Scylla: domain adaptation using frames and qualia

The methodology proposed in this paper for addressing the issue of domain adaptation in NMT systems is presented in two alternative pipelines, Scylla-S and Scylla-T, which can work around any NMT API.[2] For both of them, the most fundamental step is that of identifying the frame evoked by each LU in the sentence to be translated and, for Scylla-T, also in the translation alternatives provided by the NMT system. Next, we present how frame assignment is performed in the Scylla pipelines.

4.1. Frame disambiguation

Using frames for MT requires identifying which frames are evoked by the lexical material in the sentence to be translated. This step was implemented in the Scylla pipelines through DAISY: *Disambiguation Algorithm for Inferring the Semantics of Y*. DAISY uses the FN-Br network of typed relations as a graph, applying spread activation to estimate the frame associated with each lemma in the sentence.

Graph construction involves the following steps:

1. A dependency parser – UDPipe (Straka and Straková, 2020) – processes the input sentence for the identification of word forms and lemmas in the sentence;
2. From the lemmas obtained, the system searches for MWEs matching those in the FN-Br database;
3. Based on a simplified set of syntactic patterns, lemma clusters are defined. Each cluster contains the lemmas which are directly associated to each other;
4. LUs associated to each lemma are retrieved from the FN-Br database;
5. Qualia relations holding between LUs in a cluster are retrieved;
6. The frame evoked by each LU is retrieved form the FN-Br database, for each frame, related frames are also stored;
7. FE-to-Frame relations are retrieved so that the frame evoked by the LU can be related to other frames evoked by other LUs in the cluster.

All those elements (word forms, lemmas, LUs, and frames) are used as nodes in the graph built and the relations between them are the links connecting the nodes.

DAISY uses the spread activation search method to traverse the graph. The process is initiated by posing an "energy level" or "activation" to a set of initial nodes and then iteratively propagating that activation out to other nodes linked to the source nodes. Every time a link is traversed, the activation values decay according to a predefined formula. If a target node receives activation from more than one source node, its activation value is incremented. This method allows for measuring the relative importance of each node in the network, considering not only how much the node is distant from the initial nodes, but also how it connects to other nodes.

Spread activation for DAISY uses real activation values. The initial nodes receive an activation value of

[1]The FN-Br database currently features 41 types of ternary qualia relations (TQRs), which are listed in Appendix A.

[2]For the implementations reported in this paper the NMT system used is the Google Translate V2 API.

1.0. Each link type is assigned a weight. This weight is used to decrease the activation value. Current implementation applies the following weights to relations: (i) frame evocation: 1.0; (ii) frame inheritance: 1.0; (iii) frame perspective: 0.9; (iv) subframe: 0.8; (v) frame element to frame: 0.5; (vi) qualia relations: 0.9. For each iteration p, a given node j has an activation level represented by $A_k(p)$ and generates an output $O_j(p)$, being that a function of its activation level, according to Equation 1.

$$A_k(p) = \sum_j O_j(p-1)W_{jk} \quad (1)$$

The output function $O_j(p)$, a variation of the logistic function represented in 2, was carefully chosen to avoid excessive activation of the nodes in the network.

$$O_j(p) = \frac{1 - exp(5 * (-A_j(p)))}{1 + exp(-A_j(p))} \quad (2)$$

The spread activation process occurs until it reaches the frames. At this point, a backpropagation process is initiated, applying the activation level calculation again at each node until reaching the LUs. Hence, each LU is assigned a weight indicating its relative importance in the network. Finally, the relative weight of each LU is associated to each lemma. This is made by adding up the activation level of each LU and dividing the result by the number of related LUs. The frame evoked by the LU with the highest relative weight is taken as the frame associated with the lemma.

The frame assignment and/or disambiguation process performed by DAISY has the advantage of not depending on large annotated datasets for training, as it is the case of SEMAFOR (Chen et al., 2010) and Open Sesame (Swayamdipta et al., 2017). Moreover, because it takes the sentence context into consideration when assigning the best fit frame for polysemous lemmas, it helps avoid performance loss for out of domain sentences. Consider, for example, the br-pt sentences in (1) and (2). Note that *bandeja.n* is a polysemous lemma in br-pt. While in (1) it evokes the `Winning_moves` frame, in (2), it evokes `Utensils`. DAISY is capable of correctly identifying each of those frames because of the surrounding context in each sentence.[3]

(1) O jogador de basquete converteu a **bandeja**.
The basketball player scored the lay-up.

(2) O garçom colocou as tijelas na **bandeja**.
The waiter put the bowls on the tray.

Both versions of Scylla use DAISY to acquire information on the frames evoked by the source sentence and, for the Scylla-T version, also by the target sentence. Next, we present each version of the pipeline in detail.

[3]The frame assignment graphs generated by DAISY for sentences in (1) and (2) are given in Appendix B.

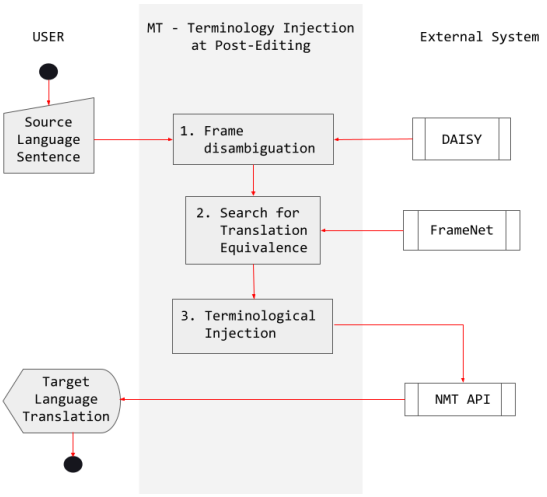

Figure 2: The Scylla-S pipeline

4.2. Scylla-S: terminology injection during the pre-processing stage

In Scylla-S, the process of terminology injection occurs in a pre-processing stage. The source sentence is submitted to DAISY and Scylla-S substitutes the source words or MWEs for which it found an entry in the FN-Br database for their translation equivalents. The hybrid sentence is then submitted to the NMT system, which, in turn, is set so that it copies unknown words into the target sentence. Because most translations are not part of the set of known words in the source language, the resulting sentence usually contains the domain-adequate expression.

For example, when the source sentence in (3) is submitted to Scylla-S, it generates the hybrid sentence in (4). This hybrid sentence is then submitted to the NMT API, yielding (5) as an output. A summary of Scylla-S is presented in Figure 2.

(3) O ponta é o jogador que menos tempo tem para pensar na armação de uma jogada.
The wing is the player with less time think about setting up a play

(4) O **wing** é o **player** que menos tempo tem para pensar na armação de uma **play**.

(5) The wing is the player **that has less time to think in the setup of** a play.

Scylla-S has some limitations. First, because the sentence fed into the NMT system is a hybrid – (4) –, containing words in both the source and the target languages, the performance of the NMT system decays sensibly, mainly regarding fluency – see boldface fragment in (5). Moreover, sometimes the target language word happens to be found in the source language vocabulary of the NMT system. For instance, one of the sentences in the experimental dataset had the LU *levantamento.n*, which, in the domain of weightlifting, is equivalent to *lift.n*. However, when the hybrid sentence

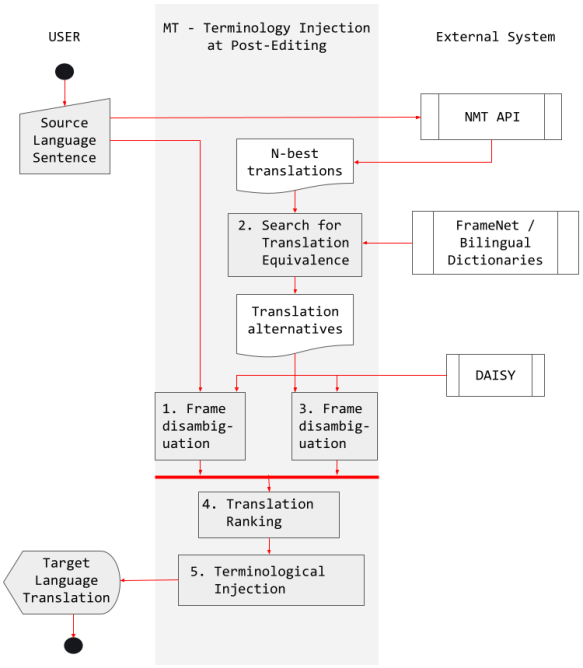

Figure 3: The Scylla-T pipeline

is fed to the NMT API, it recognizes *lift.n* as a br-pt word used in the domain of plastic surgery and, therefore, translates it again into *facelift.n*.

In an attempt to overcome the limitations of Scylla-S, we developed Scylla-T.

4.3. Scylla-T: terminology injection during the post-editing stage

Scylla-T performs terminology injection in the target sentence. For this process to work properly, it is necessary to align the words and MWEs in the source sentence with those in the sentence generated by the NMT API. First, the source sentence is submitted to the NMT API and n-best translations are retrieved. Next, Scylla-T queries the FN-Br database and also a bilingual dictionary API[4] for all possible translation equivalents of the words and MWEs in the n-best translations generated by the NMT API. The bilingual dictionary is used as means of guaranteeing that words not included in the FN-Br database are also covered by the process.

The translation equivalents retrieved are compared to the original words or MWEs in the source sentence and their synonyms in the source language. Whenever the system finds a match, an alignment is created for a word pair. The Jaro-Winkler similarity metric is used so that different word forms are considered when generating the matches.[5] Once the alignment is concluded, equivalence sets are concatenated to compose the set of translation alternatives for each word in the target sentence.

Such an approach is based on a binary tree search, where each node represents a possible point for terminology injection at the target sentence. When building the tree, nodes are created for each translation alternative. To avoid node duplicity and to assure that no translation alternative is ignored, tree construction is performed recursively, so that every possible combination of terminology injection for one same word or MWE is considered.

Because the binary tree search aims to find the translation alternative with the highest semantic similarity with the source sentence, the definition of an objective function is necessary. We defined a semantic similarity metric based on the frames evoked by each sentence, which were extracted using DAISY. Equation 3 computes the semantic similarity between two sentences from the number of coincidental frames found in both. F_o and F_d are the set of frames extracted by DAISY for the source and target sentences, respectively. The binary tree search results in the maximization of Equation 3 and is used as a heuristics for terminology injection in the target sentence, either by re-ranking one of the n-best sentence translation alternatives, or by the substitution of equivalents that are not domain-compliant by in-domain terms form the FN-Br database. Therefore, the final output of Scylla-T is a translation with a higher semantic similarity with the source sentence.

$$S(F_o, F_d) = \sum_{f_o \in F_o}^{|F_o|} \sum_{f_d \in F_d}^{|F_d|} 1[f_o = f_d] \qquad (3)$$

Scylla-T improves the performance of Scylla-S in three ways. First, because it avoids the submission of hybrid sentences to the NMT API. Second, because it may operate by just re-ranking one of the n-best sentence translation alternatives generated by the NMT, preserving its performance in which concerns fluency. Third, because, if it needs to substitute a term by another one that is domain-compliant, it does it in a more localized and precise fashion, avoiding performance loss caused by terminology injection. Those improvements can be exemplified in the translation generated by Scylla-T to the source sentence in (3), given in (6). When compared to the translation generated by the NMT API alone, given in (7), (6) is equally fluent, but terminologically accurate, since *ponta.n* in br-pt is to be translated as *wing.n* or *winger.n* in en, not as *forward.n*.

(6) The winger is the player who has less time **to think about setting up** a play.

(7) The forward is the player who has less time to think about setting up a move.

5. Evaluation

To evaluate the performance of Scylla-S and Scylla-T, a sentence translation experiment for the br-pt–en lan-

[4]For the implementation reported in this paper, the Oxford Dictionary API was used under a free academic license.

[5]Jaro-Winkler was preferred over Levenshtein because the first assigns higher values to word pairs that are more similar towards the left frontier. Because we wanted to preserve the performance of the NMT API in correctly inflecting words in the target language, Jaro-Winkler was chosen.

guage pair was designed. Both pipelines were evaluated against a commercial NMT API, namely the Google Translate V2 API, used as a baseline for BLEU, TER, and HTER. Dataset, experimental setup and results are presented next.

5.1. Dataset

The dataset used was specifically assembled for the experiment reported in this paper. It is composed by a set of sentences of the Sports domain in br-pt and a reference translation for each of them in en.[6]

Source language sentences: For conducting the translation experiment, a dataset containing 50 br-pt sentences from the Sports domain was created. All sentences were extracted from naturally produced texts, that is, from instances of textual genres such as news, encyclopedias, blog posts etc. produced by native speakers of br-pt and published in newspapers, magazines, books, websites, and the like. To make sure that the performance of the systems evaluated for domain adaptation is actually measured, for each sentence, there was at least one polysemous lemma, with at least two possible meanings, one of which related to the Sports domain. The average number of polysemous lemmas per sentence was 2.16, and the average number of possible meanings per polysemous lemma was 2.17.

Reference translation: The 50 br-pt sentences were then translated into English by a professional translator who is a native speaker of English. The translated sentences were then revised for morphosyntactic aspects and fluency by four native speakers of English. Next, sentences were analyzed for frame preservation in the Sports domain, following the Primacy of the Frame model of translation (Czulo, 2017). This is to say that, for each sentence, one linguist checked whether the frames evoked by the source and the target sentences were the same, regarding the Sports domain. The percentage of frame preservation was 72.4% of all Sports frames evoked. Provided that it is expected that translations may reorganize sentence structure, suppressing words or substituting them by hypernyms, for example, such a percentage is indicative of high semantic similarity. The reference translation sentences were used as a gold standard for evaluating the performance of the proposed solutions, as described next.

5.2. Experiments

To evaluate the performance of Scylla-S and Scylla-T, the 50 sentences in br-pt were submitted to the pipelines described in 4.2 and 4.3. Sentences are also submitted to NMT API used in Scylla-S and Scylla-T, which is considered as the baseline for comparison. Experiments were performed on a Ubuntu 20 system with 20GB RAM and a 3.1 GHz Intel i5 7200U processor.[7]

Performance was evaluated for BLEU (Papineni et al., 2002), TER, and HTER (Snover et al., 2006). BLEU evaluates correspondences between n-grams in the translations produced by each system and the gold standard translations. According to the interpretation criteria for this metric, the higher the score, the better the translation. This metric does not focus on evaluating the preservation of the semantics of the source sentence in the target sentence. Because our aim is to evaluate the performance of Scylla-S and Scylla-T for domain adaptation in NMT, we chose two other metrics – TER and HTER – that measure the effort required in post-editing a sentence generated by a MT system so that it can be regarded as a fluent, adequate translation of the source sentence.

TER uses the gold standard translations as reference to compute the minimal number of edits – additions, deletions, substitutions – that would be required in the sentence produced by the MT system so that it matches the reference sentence. Because TER is unable to assess the semantics of the MT sentence in contrast with that of the gold standard, it may end up proposing edits that are not needed, since the MT sentence just presents a different, but yet fluent and adequate, translation of the source sentence.

To overcome this limitation, HTER uses expert human translators to perform the edits so that the MT sentence becomes fluent and semantically similar to the gold standard. To avoid a high influence of human subjective choices while editing the MT sentences taking the gold standard sentences as a reference, three professional translators were hired for the task. Each of the three professional translators independently edited each of the 150 machine translated sentences (50 from each system used) taking the gold standard translations as a reference. The calculation of HTER was then based on the average number of edits proposed for each sentence, given the number of edits made by each professional translator. The computation of the number and types of edits was also carried out independently by two reviewers and revised by a third person.

Since TER and HTER measure the effort required for editing the MT sentence, the lower the score, the better the translation.

5.3. Results and Discussion

The performance of the three systems evaluated in the experiment for the three metrics chosen is presented in Table 2.

The performance of Scylla-S is the worst among the three systems for BLEU. This may be due to the fact that Scylla-S feeds hybrid sentences to the NMT API, compromising its performance at the formal pole of language structures. Even considering that BLEU does

[6] All sentences in the dataset are available at https://github.com/FrameNetBrasil/scylla_lr. See also section 7.

[7] The baseline system used can be accessed at https://cloud.google.com/translate.

	Baseline	Scylla-S	Scylla-T
BLEU	53.13	48.12	**53.66**
TER	**36.23**	42.63	36.47
HTER	13.80	10.44	**7.38**

Table 2: Evaluation of the baseline, Scylla-S and Scylla-T systems for BLUE, TER and HTER

not evaluate semantic adequacy directly, the Baseline and Scylla-T have a similar performance for this metric.

The morphosyntactic problems caused by sentence hybridization in Scylla-S are also the reason for a poorer performance of this system when evaluated for TER. Because TER is applied automatically and computes all edits needed to turn the MT output into an exact copy of the reference translation, the gains that derive from the fact that sentences produced by Scylla-S are terminologically adequate to the domain are surpassed by the losses in fluency. Once again, the Baseline and Scylla-T have a superior and very similar performance for this metric. This demonstrates that Scylla-T is able to mitigate the problems caused by the fact that Scylla-S feeds hybrid sentences to the NMT API.

For HTER, both Scylla-S and Scylla-T outperform the Baseline system. This is due to the fact that, when professional human translators are asked to adequate the MT sentence using the domain-specific gold standard as a reference, they take into consideration the fact that polysemous lemmas in the source language may have different translations in the target language, depending on the domain. Because sentences were extracted from in-domain real texts, the context provided by the sentences was able to indicate that the translation should be adequate to the Sports domain. Moreover, human translators are able to consider a sentence containing, for instance, terms that are synonymous to the ones in gold standard as equally adequate to the domain. Therefore, the performance of the systems measured by HTER is considerably better than that measured by TER.

Given the nature of the problem tackled in this paper, that of domain adaptation in NMT, we claim that HTER, although having a higher cost, is the most reliable metric for assessing the performance of the three systems in the task proposed. To illustrate, consider once again the examples already discussed in sections 4.2 and 4.3, presented together with the gold standard translation in (8-12).

(8) O ponta é o jogador que menos tempo tem para pensar na armação de uma jogada.
Source sentence

(9) The winger is the player with less time to think about setting up a strike.
Gold standard translation

(10) The **forward** is the player who has less time to think about setting up a move.
Baseline (TER=26.66 / HTER=0.08)

(11) The wing is the player that has less time **to think in the setup of** a play.
Scylla-S (TER=53.33 / HTER=0.06)

(12) The winger is the player who has less time to think about setting up a play.
Scylla-T (TER=20.00 TER / HTER=0.00)

Note that the translation generated by the Baseline System does not feature an adequate translation for *ponta.n* in the domain, while the ones generated by Scylla-S and Scylla-T do. However, the translation by Scylla-S is less fluent, while the one by Scylla-T is not, leading to lower TER and HTER values.

From the performance results for HTER in Table 2, we conclude that the domain adapted translations generated by both Scylla-S and Scylla-T present better semantic correlation with the gold standard. Scylla-T was also able to address the limitations of Scylla-S, improving the performance of the Baseline NMT system by almost 47%.

6. Conclusions and Outlook

In this paper, we presented two systems for domain adaptation in NMT using terminology injection based on qualia-enriched FrameNet. Neither system require fine-tuning of the NMT model, reducing computational costs, and mitigating performance losses due to overfitting. Evaluation of the systems' performance against a NMT API taken as baseline for the BLEU, TER, and HTER metrics demonstrated that one of the systems, Scylla-T, meets the performance of the baseline for both BLEU and TER. For the HTER metric, more adequate for measuring semantic adequacy, both systems proposed outperform the baseline, Scylla-T improving it by 47%.

Scylla-S and Scylla-T are the first systems to our knowledge to leverage framenet data and qualia relations for domain adaptation in NMT. Also, because they are both implemented as pipelines, they can be easily adapted to any NMT API.

For future work, we plan to explore the automatic expansion of domain-specific framenet coverage for other languages via the acquistion of LUs for languages still not covered by FN-Br from other lexical resources such as multilingual WordNets (Fellbaum and Vossen, 2012), or BabelNet (Navigli et al., 2021). The idea is that, because the FN-Br database already models equivalences between domain-specific lexical items in br-pt and en, those equivalences could be used as proxies for extracting equivalent terminology in other languages, making Scylla available in other languages without the need to rebuild the domain model from scratch.

7. Ethics Statement

Both the gold standard translations and the evaluation based on the HTER metric were carried out by professional translators working for a translation company

hired for that purpose. Translators worked under standard working contract regulations. Experiments were conducted without any major computational costs, as it can be inferred from the specifications of the computational environment provided in 5.2.

8. Acknowledgements

The FrameNet Brasil lab is funded by CAPES PROBRAL grant 88887.144043/2017-00. Diniz da Costa's research was funded by CAPES PROBRAL PhD exchange grant 88887.185051/2018-00. Marim's research internship was funded by CNPq BDTI program. The development of Scylla was partially funded by FAPEMIG Universal Grant APQ-00471-15. Authors thank Oliver Czulo and Alexander Ziem for their contribution to the development of this research project.

9. Bibliographical References

Arcan, M. and Buitelaar, P. (2017). Translating domain-specific expressions in knowledge bases with neural machine translation. *CoRR*, abs/1709.02184.

Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas, November. Association for Computational Linguistics.

Baker, C. F. and Lorenzi, A. (2020). Exploring crosslinguistic frame alignment. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 77–84, Marseille, France, May. European Language Resources Association.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November. Association for Computational Linguistics.

Chen, D., Schneider, N., Das, D., and Smith, N. A. (2010). SEMAFOR: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden, July. Association for Computational Linguistics.

Chu, C. and Wang, R. (2018). A survey of domain adaptation for neural machine translation. *CoRR*, abs/1806.00258.

Costa, A. D. and Torrent, T. T. (2017). A Modelagem Computacional do Domínio dos Esportes na FrameNet Brasil. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 201 – 208. Sociedade Brasileira de Computação.

Costa, A. D., Gamonal, M., Paiva, V., Marção, N., Peron-Corrêa, S., Almeida, V., Matos, E., and Torrent, T. (2018). FrameNet-Based Modeling of the Domains of Tourism and Sports for the Development of a Personal Travel Assistant Application. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 6 – 12. European Language Resources Association.

Costa, A. D. (2020). *A tradução por máquina enriquecida semanticamente com frames e papéis qualia*. Ph.D. thesis, Universidade Federal de Juiz de Fora.

Czulo, O. (2017). Aspects of a primacy of frame model of translation. In Silvia Hansen-Schirra, et al., editors, *Empirical modelling of translation and interpreting*, pages 465–490. Language Science Press, Berlin.

Dougal, D. K. and Lonsdale, D. (2020). Improving NMT quality using terminology injection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4820–4827, Marseille, France, May. European Language Resources Association.

Fellbaum, C. and Vossen, P. (2012). Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2):313–326.

Fillmore, C. J. and Baker, C. (2009). A frames approach to semantic analysis. In Bernd Heine et al., editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK, December.

Fillmore, C. J. (1982). Frame Semantics. In Linguistics Society of Korea, editor, *Linguitics in the morning calm*. Hanshin Publishing Co., Seoul, South Korea. pages: 111 – 138.

Khayrallah, H., Thompson, B., Duh, K., and Koehn, P. (2018). Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages

36–44, Melbourne, Australia, July. Association for Computational Linguistics.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., et al. (2000). Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

Moussallem, D., Ngonga Ngomo, A.-C., Buitelaar, P., and Arcan, M. (2019). Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 139–146.

Navigli, R., Bevilacqua, M., Conia, S., Montagnini, D., and Cecconi, F. (2021). Ten years of babelnet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pustejovsky, J., Havasi, C., Littman, J., Rumshisky, A., and Verhagen, M. (2006). Towards a generative lexical resource: The Brandeis semantic ontology. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press, Cambridge, Mass.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Straka, M. and Straková, J. (2020). UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 124–129, Marseille, France, May. European Language Resources Association (ELRA).

Swayamdipta, S., Thomson, S., Dyer, C., and Smith, N. A. (2017). Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Thompson, B., Gwinnup, J., Khayrallah, H., Duh, K., and Koehn, P. (2019). Overcoming catastrophic forgetting during domain adaptation of neural machine translation. 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 2062–2068, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Torrent, T. T., Salomao, M. M. M., da Silva Matos, E. E., Gamonal, M. A., Gonçalves, J., de Souza, B. P., Gomes, D. S., and Peron-Corrêa, S. R. (2014). Multilingual lexicographic annotation for domain-specific electronic dictionaries: The copa 2014 framenet brasil project. *Constructions and Frames*, 6(1):73–91.

Torrent, T. T., da Silva Matos, E. E., Lage, L. M., Laviola, A., da Silva Tavares, T., de Almeida, V. G., and Sigiliano, N. S. (2018). Towards continuity between the lexicon and the constructicon in FrameNet Brasil. In Benjamin Lyngfelt, et al., editors, *Constructicography: Constructicon development across languages*, pages 107–140. John Benjamins, Amsterdam.

Zhang, J. and Zong, C. (2016). Bridging neural machine translation and bilingual dictionaries. *CoRR*, abs/1610.07272.

Appendices

A – Ternary Qualia Relations in the FN-Br Database

The distribution of ternary qualia relations (TQRs) per quale and per language is shown in Table 1.

Table 3 presents the 41 TQRs modeled in FN-Br. The first column shows the quale refined by the TQR, while the second gives a mnemonic key for the relation. The third column shows the frame mediating the TQR, while the last two show the core FEs which can be prototypically instantiated by the two LUs involved in the TQR.

B – Example frame assignment graphs generated by DAISY.

Figures 4 and 5 depict the graphs generated by DAISY for assigning the correct frames to *bandeja.n*, according to the context provided by each sentence. Both figures were generated

Quale	Relation	Frame	FE1	FE2
Agentive	created by	Achieving_first	NEW_IDEA	COGNIZER
Agentive	caused by	Causation	EFFECT	Actor
Agentive	caused by	Causation	EFFECT	Cause
Agentive	created by	Cooking_creation	PRODUCED_FOOD	COOK
Agentive	caused by	Intentionally_act	ACT	AGENT
Agentive	affected by	Intentionally_affect	PATIENT	AGENT
Agentive	created by	Intentionally_create	CREATED_ENT.	CREATOR
Agentive	resolved by	Resolve_problem	PROBLEM	AGENT
Constitutive	has as attribute	Attributes	ENTITY	ATTRIBUTE
Constitutive	has as part	Building_parts	WHOLE	PART
Constitutive	causes	Causation	ACTOR	AFFECTED
Constitutive	contains	Containing	CONTAINER	CONTENTS
Constitutive	produces	Creating	CREATOR	CREATED_ENT.
Constitutive	workplace of	Employing	EMPLOYER	EMPLOYEE
Constitutive	includes	Inclusion	TOTAL	PART
Constitutive	used by	Infrastructure	INFRASTRUCTURE	USER
Constitutive	made of	Ingredients	PRODUCT	MATERIAL
Constitutive	performed by	Intentionally_act	ACT	AGENT
Constitutive	relative of	Kinship	EGO	ALTER
Constitutive	has as member	Membership	GROUP	MEMBER
Constitutive	affects	Obj_influence	INFLUENCING_ENT.	DEPENDENT_ENT.
Constitutive	has as part	Part_inner_outer	WHOLE	PART
Constitutive	has as part	Part_piece	SUBSTANCE	PIECE
Constitutive	has as part	Part_whole	WHOLE	PART
Constitutive	has origin at	People_origin	PERSON	ORIGIN
Constitutive	follower of	People_religion	PERSON	RELIGION
Constitutive	relates to	Relation	ENTITY_1	ENTITY_2
Constitutive	has as resident	Residence	LOCATION	RESIDENT
Constitutive	uses	Using_resource	AGENT	RESOURCE
Formal	instance of	Exemplar	INSTANCE	TYPE
Formal	type of	Type	SUBTYPE	CATEGORY
Telic	vice of	Addiction	ADDICTANT	ADDICT
Telic	ability of	Capability	EVENT	ENTITY
Telic	habit of	Custom	BEHAVIOR	PROTAGONIST
Telic	performed at	Infrastructure	ACTIVITY	INFRASTRUCTURE
Telic	activity of	Intentionally_act	ACT	AGENT
Telic	created by	Intentionally_create	CREATED_ENT.	CREATOR
Telic	purpose of	Purpose	GOAL	AGENT
Telic	purpose of	Tool_purpose	PURPOSE	TOOL
Telic	used for	Using	AGENT	PURPOSE
Telic	used by	Using_resource	RESOURCE	AGENT

Table 3: Ternary Qualia relations in FN-Br.

using the DAISY demo interface available at http://server3.framenetbr.ufjf.br:8010/index.php/daisy/main.

Note, in Figure 4, that DAISY is able to properly disambiguate the lemma *bandeja.n*, shown in cluster 507. In the FN-Br lexicon, this lemma may evoke three different frames: Artifact and Utensils, being equivalent to *tray.n* and Winning_moves, being equivalent to *lay up.n*. In the context of a sentence where the subject is a basketball player, the last frame is the correct one. DAISY reaches an activation level of 4.01 for the Winning_moves frame, against a 0.53 level for the other two. DAISY also correctly disambiguates between the two senses of *converter.v* – cluster 505 –, assigning an activation level of 4.18 to the Winning_moves frame, against only 0.50 to the Undergo_transformation frame. In en, the equivalent LUs evoking those frames would be *score.v* and *turn into.v*, respectively. Finally, it is worth noting that DAISY correctly identifies the MWE *jogador de basquete.n*, equivalent to *basketball player.n*, as shown in clusters 502, 503 and 504. In the first, the MWE is preferred over *jogador.n* alone. In the third, over *basquete.n* alone.

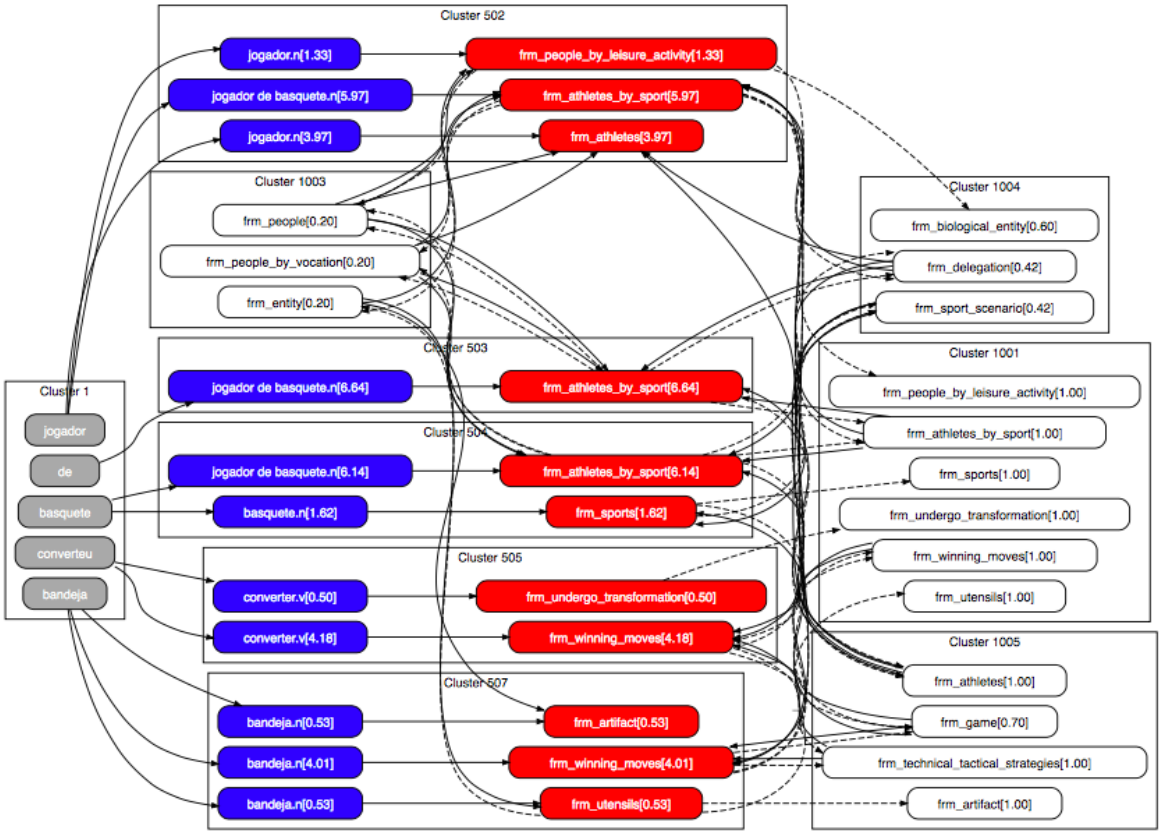

Figure 4: The frame assignment graph generated by DAISY for example sentence (1).

A similar scenario holds for Figure 5, where DAISY correctly identifies *bandeja.n* as an LU evoking the `Utensils` frame with ans activation level of 3.13.

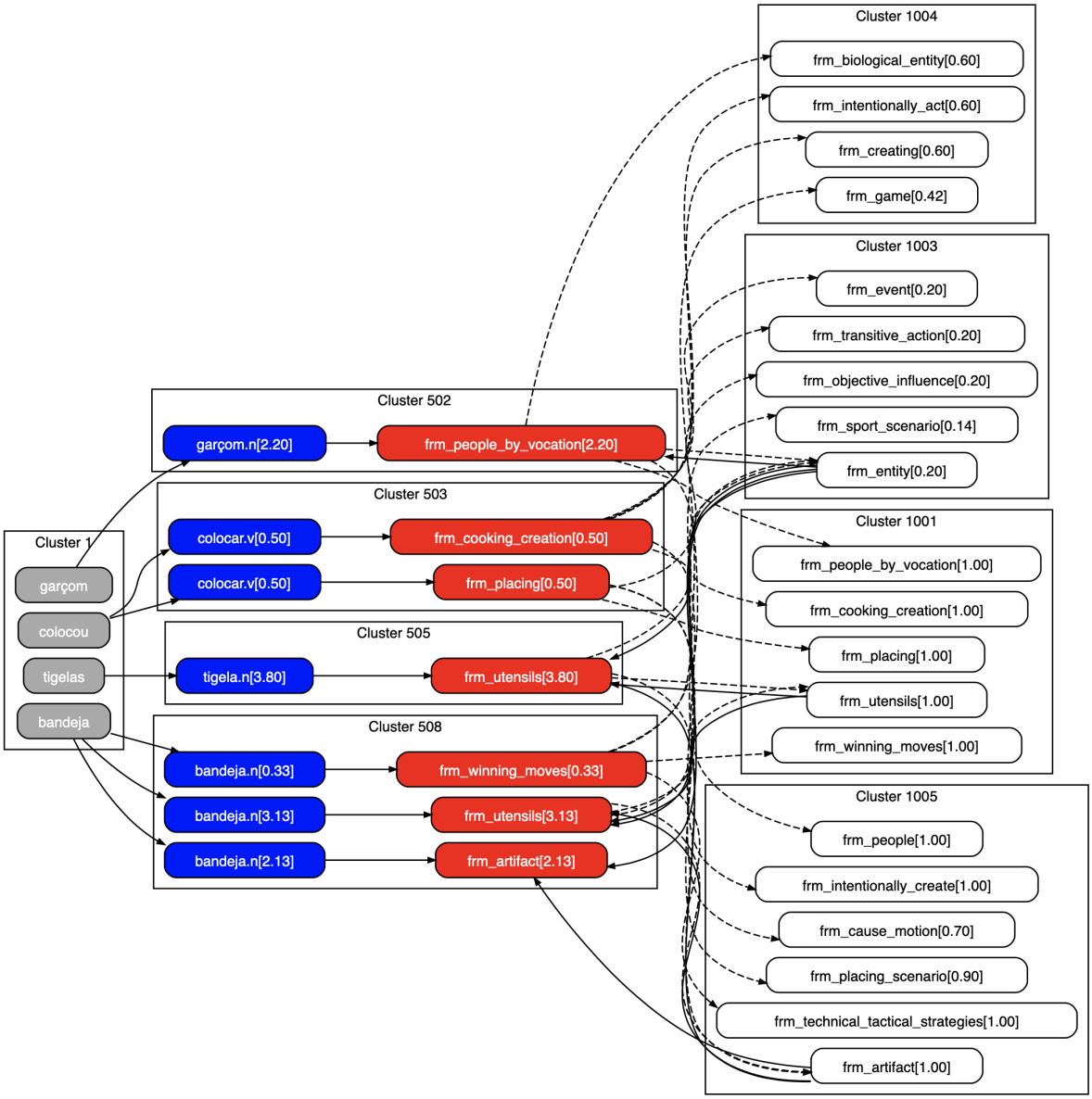

Figure 5: The frame assignment graph generated by DAISY for example sentence (2).