Automatic Identification of Equivalence of Concepts in Different Languages for Never-Ending Learning

Silvio C. Marino[1], Estevam R. Hruschka Junior[2]

[1]Departamento de Computação
Universidade Federal de São Carlos (UFSCar) – São Carlos, SP – Brazil

[2]Departamento de Computação
Universidade Federal de São Carlos (UFSCar) – São Carlos, SP – Brazil

silviomarino@gmail.com, estevam.hruschka@gmail.com

Abstract. *This paper describes the process of automatic identification of concepts in different languages using a base that relies on simple semantic and morphosyntactic characteristics like string similarity, difference in words amount and translation position on dictionary (when exists) and a neural network that has been used as a model of machine learning. All experiments use data that was obtained from a few categories of Read The Web (RTW) project and an endless learning computation system called NELL: Never-Ending Language Learning. The results were compared with dictionary and showed that the introduction of neural network brought a significant gain in the process of equivalence of concepts.*

1. Introduction

When we speak of Artificial Intelligence (AI) we refer to a machine capable of systematizing and automating tasks "that require intelligence when performed by people" [Kurzweil 1990].

Machine Learning (ML) is one area of AI that seeks the development of computer programs that can evolve as they are exposed to new experiences [Mitchell 1997]. The main objective of ML is to search for methods and techniques that allow the design of computational systems capable of improving its performance, autonomously and based on information obtained through its use; which is considered one of the fundamental mechanisms governing automatic learning processes [Mitchell 2006]. With ML it may be possible to create a decision-making process with an ever-smaller margin of error.

Machine learning can be divided into three types: supervised, unsupervised and semi-supervised learning [Zhu et al., 2003]. Supervised learning consists of a task of creating a model obtained by training the model from the known data set and its labels. The goal is to extract the rule that relates entries to labels. In unsupervised learning, no type of label is given for learning. The idea is that algorithms have the ability to form groups based on the similarity of the data. Finally, a form of semi-supervised learning is to use a little of supervised approach and a bit of unsupervised approach. Its main feature is, from a small sample of data with labeled, extract the model (supervised learning) to label a large amount of unlabeled data (unsupervised learning).

A variety of large knowledge bases (KBs) have been constructed e.g., DBpedia, Yago and NELL. These KBs consist of an ontology that define a set of categories (e.g., *sportsTeam, City*), the relations with these categories as arguments types (e.g.,

teamPlaysInCity(*SportsTeam, City*)), KB entities which instantiate these categories (e.g., *Steelers* ∈ *SportsTeam*) and KB entity pairs which instantiate these relations (e.g., (*Steelers, Pittsburgh*) ∈ *teamPlaysInCity*). [Wijaya 2016].

The project Read the Web and an endless learning computing system called NELL (Never-Ending Language Learning) were created in 2008. NELL is a semi-supervised learning system that learns how to read web pages. The main goal of the project is to develop an endless learning system that runs 24 hours a day, seven days a week. This system is able to extract information from web pages and improve its performance every day. Although the concept of endless machine learning is not something new, NELL has virtually no competitor since the system started in 2010.

NELL [Mitchell et al 2018] is a system that has an initial (manually defined) knowledge base composed of categories and relations. Examples of categories present in the NELL knowledge base are: city, person, company, country, product, athlete, sport, sports team, etc. Some examples of relationships are: *liveIn(person, city)*, *produceProduct(company, product)*, etc.

Based on the categories and relationships of the initial knowledge base and on a set of instances (between 10 and 20 instances) for each category and each relation, NELL realizes its endless learning process, in order to learn to read the web better and to be able to extract more facts each time with more precision.

In 2009, began the project to read the web in Portuguese with the Read The Web in Portuguese (RTWP) system. RTWP is a system created based on NELL, which learns named entities and textual standards from web pages. According to results presented in [Duarte 2011], with the first version of RTWP it was shown, through empirical evidence, that reading of the web in Portuguese is viable.

With success of obtaining knowledge in English and studies demonstrating the feasibility of reading the web in other languages, some changes were made to allow to extract facts from web pages also in Portuguese, French and Spanish. However, there is no relationship between the knowledge learned among languages.

When we think of a person who speaks several languages, it is logical to imagine that the knowledge acquired in one language can be used in any language. For example, when we read a tutorial in English about making an apple pie, we have learned to perform this activity in any language we master. This is because we have a unique way of storing and relating acquired knowledge.

Again, thinking about a human being, when we know the concept associated with a word in English and we want to know the equivalent concept in another language, we use a dictionary. Although the use of a dictionary is quite useful, there are some problems.

The first problem occurs when the dictionary does not have the word or expression that we want to know the equivalent concept. For example, we hardly find in a dictionary the expression “City that never sleeps” equivalent to “New York”.

Another situation occurs when the dictionary provides a very extensive list of equivalent concepts and it is not possible to identify which one is the most appropriate term to represent the concept of interest.

In this way, we introduce here an approach which utilize NELL’s knowledge base to generate databases supported on semantic and morphosyntactic characteristics. These

databases, are used by a machine learning model, capable of automatically saying, if with two concepts (En, Pt), En in English and Pt in Portuguese, it can be said that En is equivalent to Pt.

2. Method

In [González, Hruschka, Mitchell 2017], authors describe how they dealt with the problem of identifying equivalent concepts, in Portuguese and English, assuming that they share the same ontology of categories and relations. In the study, two knowledge bases, learned independently from different web pages written respectively in English and in Portuguese, were used. A custom PageRank approach (which can be considered as the measure of similarity that characterizes the neighborhood of an X node in a graph) and an inference technique to find common relevant paths through the knowledge bases, were used. It was found that, the proposed inference technique efficiently identifies relevant paths with better results than simple use of dictionaries, in most of categories tested.

This is one of the works that, helped to realize the need to generate, from NELL's categories, a set of data with positive examples (where equivalence is true) and a set of negative data (where equivalence is false). To facilitate future comparison between the methods used to deal with identification of equivalence of concepts, we used the same positive and negative examples that [González, Hruschka, Mitchell 2017] had used. It should be noted that only a few categories were used to generate the examples for training the learning model: actor, animal, city, country, movie, person, sport and writer. For each positive example, it was manually generated 2 negatives examples. Table 1 presents the number of positives and negatives examples of each category.

Table 1. Positive and negative examples.

Category	Number of Positive Examples	Number of Negative Examples
Actor	510	1020
Animal	60	120
City	510	1020
Country	110	220
Movie	40	80
Person	1560	3120
Sport	170	340
Writer	60	120
Total	3020	6040

For generation of English-Portuguese dictionary, it was first necessary to export existing data in NELL's knowledge base in English (the complete file contains over 2.3 million lines). After, an algorithm was developed to process all words and expressions and to remove repetitions. After that, a code was generated to use the Wikipedia and Wordnet APIs to try to translate each word or expression from English to Portuguese. As a result, we had a dictionary with 78,650 entries that have one or more translations.

When we analyze the knowledge acquired in a given language, it is possible to find cases where the object, which represents a learned information, ends up being named in several ways. These cases, are called correferentes and are very relevant to endless learning. In [Duarte, Hruschka 2014], authors describe how semantic and morphological characteristics (among them, difference in number of words and similarity of strings) were combined in resolution of the problem of coreference in NELL. Empirical evidence has been obtained to show that, combining morphological and semantic features in a hybrid model, can positively impact NELL knowledge base.

The set of attributes related to morphosyntactic characteristics, defined based on related works, like [Duarte, Hruschka 2014], can be described as an inspiration of what a person would do when given the task of saying whether a set of English words and their possible translation into Portuguese, is true or false.

In this activity, it is normal for a person to try to extract some rules based, for example, on difference in the number of words between the two terms, dictionary query, number of words that begin with capital letter and similarity between the terms.

The difference in number of words between the two terms is a valid factor that can be used to make the decision. Overall, we hope that if an English term has a single word, its translation into Portuguese will not have a much larger number of words. In addition, this task will be easier if a dictionary can be queried. If the Portuguese term appears in the translation list, there is more security to say that the translation is true.

The number of words that begin with a capital letter may also be one of the factors to consider. Names of people and places are expected to have a very close number of words beginning with capital letters. On NELL's knowledge base, all words were recorded with lowercase letters. So, the difference in the number of words beginning with a capital letter was neglected.

Finally, the similarity between terms can be compared. There are several English words that have the same or very similar spelling when translated into Portuguese. This occurs not only for names of people and places, but also for animals, fruits, objects, among others. Pairs ("banana", "banana") and ("animal", "animal") are examples of identical spellings. Tuples ("actor", "ator"), ("cat", "gato"), ("person", "pessoa") and ("cell", "cellular") are examples of similar spellings.

In this way, the morphosyntactic base has the following attributes: "En", "Pt", "Dif Num Words En - Pt", "Dictionary Position", "Similarity", "Translation". Table 2 presents a description of each of the attributes.

Table 2. Attributes of morphosyntactic base.

Attribute	Description
En	Term in English
Pt	Possible translation into Portuguese.
Dif Num Words En - Pt	Difference in the number of words in English and their possible translation into Portuguese.
Dictionary Position	Numerical value that informs the order that the Portuguese word appears in dictionary list. If the word is not found, -1.
Similarity	Value obtained from an algorithm that compares similarity between sentences, float value (0 to 1) of similarity.
Translation	For positive example value is 1, negative 0.

An algorithm was developed to performs the filling of each attribute. For each line read in the positive examples file, two lines are read in the negative file. With this, it is avoided to create blocks with only one type of example.

For morphosyntactic base, initially the English term, possible translation to Portuguese and value that informs whether the equivalence of concept is true or false are respectively stored in "En", "Pt" and "Translation" attributes. Following, it is counted the number of words in the concept in English and the number of words in the concept in Portuguese. The difference between them is stored in the attribute "Dif Num Words En - Pt".

If the concept in Portuguese exists in the dictionary as a translation of the concept in English, the position it appears in is saved in the "Position in the Dictionary" attribute. Otherwise, the attribute is -1. Finally, the value returned by invoking the method that checks for similarity between two strings is inserted in the "Similarity" attribute.

Part of the database with morphosyntactic characteristics are presented in Table 3.

Table 3. Data of morphosyntactic base.

En	Pt	Dif Num Words En - Pt	Position in the Dictionary	Similarity	Translation
goose	ganso	0	4	0.4	1
cat	golfinho	0	-1	0.0	0
tortoises	iguanas	0	-1	0.25	0
pumas	pumas	0	1	1.0	1
pests	gato	0	-1	0.43	0
insects	rato	0	-1	0.25	0

The set of attributes based on semantic characteristics, was defined based on observation of the domain in which a word or expression is inserted. For example, to the concept “cats” in English, it is expected that some of the following terms appear in domain: “pet”, “disease”, “furry”, “mouse”, among others. By logic, it is expected that something similar happens with Portuguese term, “gatos”.

Each NELL category has a specific set of relationships in its domain. For example, the category animal has some relations like: animalDevelopeDesease, animalEatVegetal, animalEatFood, etc. Since most of the existing relations in English have their corresponding in Portuguese, the idea is to use each relation as an attribute of semantic base and try to find equivalences. Table 4 show the number of relations of each category.

Table 4. Number of relations of each category.

Category	Number of Relations
Actor	122
Animal	59
City	145
Country	130
Movie	33
Person	117
Sport	37
Writer	120

Before filling the semantic base with data, the relations of all categories in English and their equivalent relation in Portuguese, when it exists, has been inserted into the dictionary. So, we get the positive and negative examples to fill the base with semantic characteristics.

The first step involves to retrieve each relation of the category of the term in English and all relations of the category of the equivalent candidate in Portuguese. The dictionary is used to translate each relation in English to Portuguese. When a relation in English hasn’t translation to Portuguese, its value is set to 0 (because no future equivalence can be founded) and the process gets the next relation. However, when it exists in the relations of the equivalent candidate in Portuguese, the attribute receives a minimum score of 1. A relationship has one or multiple values. Thus, each value of the relation in English is obtained. The dictionary is used again to try to translate the value from English into Portuguese. When the translated value exists in the equivalent values in Portuguese, the attribute has its value incremented by 1. Table 5 illustrates part of the database with semantic characteristics.

Table 5. Data of semantic base.

En	Pt	animalDevelopeDesease	...	animalEatFood	Translation
cat	gato	5	...	4	1
cat	urso	0	...	1	0
dog	rato	1	...	0	0

The last base has morphosyntactic and semantic characteristics. It was filled by simply joining the morphosyntactic data with semantic data.

It was now necessary to choose the machine learning model. Among many authors, in [Ke, Hagiwara 2015], authors present how they used neural networks to process texts in English. A neural network with 5 hidden layers was used to extract sentences, phrases, words and concepts from the text, with the objective of answering a questionnaire equal to that applied to students. Of 495 possible points, neural network achieved an average of 276 points, while the average number of students was 263.

A neural network is a machine designed to model the way the brain performs a task. Normally the network is implemented using electronic components or simulated through programming on a digital computer. Simon Haykin (2007) offers the following definition of a neural network seen as an adaptive machine: A neural network is a massively distributed processor made up of simple processing units that have the natural propensity to store experimental knowledge and make it available for use.

Perceptron is the simplest form of a neural network used to classify patterns that lie on opposite sides of a hyperplane, that is, linearly separable. With a single neuron it is possible to classify only patterns with two hypotheses. By expanding the output layer to more neurons, it is possible to classify more classes as long as they are linearly separable. The perceptron is composed of a single neuron with synaptic weights that can be adapted from iteration to iteration and bias.

Multi-layer perceptron (MLP) networks consist of a set of sensory units that constitute the input layer, one or more hidden layers of computational nodes, and, finally, an output layer. The input signal is propagated forward through the network, layer by layer. The training takes place in a supervised way through an algorithm known as back-propagation error algorithm, which is based on the error correction learning rule. Simply put, the error backpropagation consists of two steps through the different layers of the network: one step forward, the propagation, and one step backward, the back-propagation.

Based on the literature review, a neural network was chosen as the machine learning model. The network was trained using 5-fold cross-validation. This means that the base was divided into five parts and the network was run five times. At each iteration, one part is removed to be used as a test, while the other four are used as a training set. Before executing neural network, all data were normalized, that is, mean zero and standard deviation one.

3. Experiments and Analysis

Experiments were performed on the following bases: morphosyntactic, semantic, and morphosyntactic merged with semantic. At each base, the experiments were performed with data partitioned by category, as well as, with the unified base, which brings together all categories. Table 6 presents all experiments performed.

Table 6. Experiments performed.

Base	Category	
	Partitioned	Unified
Morphosyntactic	X	X
Semantic	X	X
Morphosyntactic and Semantic	X	X

Results obtained by MLP were compared with three cases of use of the English-Portuguese dictionary. In the first case, when the English term is found in the dictionary, the comparison with the Portuguese term is made with the first available translation. In the second case, comparison with the Portuguese term is done with a random translation among all available. The last type of comparison is done by verifying the existence of the Portuguese term in the list of translations of the English term. Figure 1 presents the neural network compared with the three dictionary use cases.

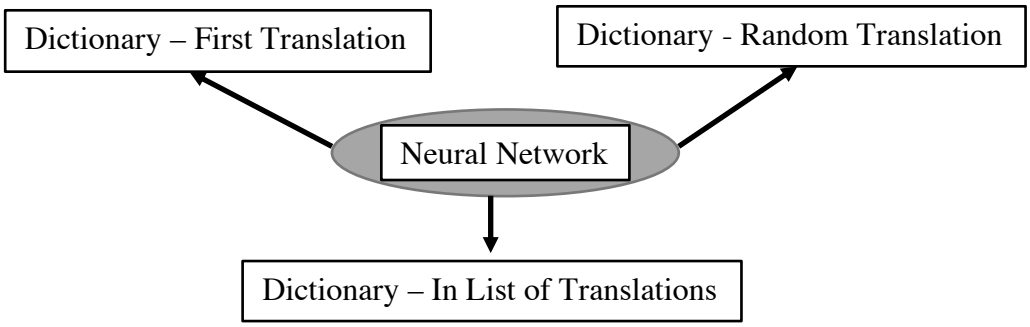

Figure 1. Neural Network compared with dictionary use cases.

All set of experiments were carried out using MLP, as machine learning model, changing the number of hidden layers (1 through 5) and the learning rate: 0.1, 0.05, 0.01 and 0.005.

3.1 Experiments with Morphosyntactic Base – Unified Categories

First set of experiments were carried out with the morphosyntactic base with all categories together. Table 7 presents the confusion matrix of the neural network, dictionary - first translation, dictionary - random translation and dictionary - in list of translations, for the best result that occurred with 3 hidden layers and a learning rate of 0.005.

Table 7 - Matrix of confusion of neural network and dictionaries in the morphosyntactic base with all categories together

Neural Network			Dictionary - First			Dictionary - Random			Dictionary - In List		
	0	1		0	1		0	1		0	1
0 (Real)	**6024**	16	0	**6039**	1	0	6038	2	0	6035	5
1 (Real)	106	**2914**	1	2356	664	1	2364	656	1	2267	**753**

Comparing results of dictionaries, for positive examples, it is possible to say that the result obtained using the three methods was similar: 6039 hits using the first translation of the dictionary; 6038 using random translation among all available and 6035 verifying if the candidate is in the list of translations of the English term. The result of MLP (6024), for negative examples, was somewhat inferior to the use of dictionaries. However, when analyzing 2914 positive examples agreed by the neural network, it is noted that the number of correct answers of the MLP was much higher than 753 correct ones using the best dictionary result which happened when the Portuguese term is in the list of all translations of the English term. MLP was more than three times better.

3.2 Experiments with Morphosyntactic Base - Partitioned by Categories

Second set of experiments were performed with morphosyntactic base partitioned by categories. Table 8 presents, for each category, neural network confusion matrix, dictionary - first translation, dictionary - random translation and dictionary – in list of translations, for the best result that occurred with 1 hidden layer and a learning rate of 0.01.

Table 8 - Matrix of confusion of the neural network and dictionaries in the morphosyntactic base partitioned by the categories.

Neural Network			Dictionary - First			Dictionary - Random			Dictionary - In List		
	0	1		0	1		0	1		0	1
0 (Real)	**6009**	31	0	**6039**	1	0	6038	2	0	6035	5
1 (Real)	80	**2940**	1	2356	664	1	2364	656	1	2267	**753**

Analyzing only the negative examples of the neural network, it is possible to verify that the amount of correctness continued with a slightly inferior result compared with dictionaries. In addition, there was a small decrease (6009) compared with the first set of experiments (6024).

On the other hand, observing only the positive examples of the neural network, it is possible to say that the amount of correctness also continued much higher than the use of dictionaries. In addition, there was a small increase (2940) compared with the first set of experiments (2914).

By adding up the number of correct positive and negative examples, it is possible to verify that these experiments produced a little higher result (8949) than the first set of experiments (8938).

3.3 Experiments with Semantic Base

As mentioned in section 2, each relation of the categories is used as attributes of the semantic base. Thus, from Table 4, it is possible to observe that the lowest semantic base, partitioned by category, has 33 attributes. And the largest has 145 attributes.

The structure of the semantic database, unified by the categories, was generated from the simple addition of the relations, removing repetitions. In this way, the semantic base, with all categories together, has 284 attributes.

All experiments, based only on semantic (unified or partitioned) features, failed. This may have happened because of the large number of zeros in attributes that prevented the neural network to converge. This indicates that only with this data, there is no semantic information capable of aiding the automatic identification process of equivalence of concepts.

3.4 Experiments with Morphosyntactic merged with Semantic Base – Unified Categories

The next set of experiments were performed on the base that brings together all the categories and combines the morphosyntactic and semantic characteristics. In total the base has 287 attributes. Table 9 presents the confusion matrix of the neural network, dictionary - first translation, dictionary - random translation and dictionary - in list of translations, for the best result that occurred with 2 hidden layers and a learning rate of 0.01.

Table 9 - Matrix of confusion of the neural network and of the dictionaries in the morphosyntactic and semantic base with all categories together.

Neural Network			Dictionary - First			Dictionary - Random			Dictionary - In List		
	0	1		0	1		0	1		0	1
0 (Real)	**6019**	34	0	**6039**	1	0	6038	2	0	6035	5
1 (Real)	166	**2895**	1	2356	664	1	2364	656	1	2267	**753**

Analyzing only the negative examples of the neural network (6019), it is possible to verify that the amount of correctness continued with a slightly inferior result compared with dictionaries. In addition, there was no gain compared to the best result for negative examples, found in the first set of experiments (6024).

Observing the positive examples of the neural network (2895), it is possible to say that the amount of correctness also continued much higher than the use of dictionaries. However, these experiments did not outperform the best result, for positive examples, found in the second set of experiments (2940).

By adding up the number of correct positive and negative examples (8914), it is verified that these set of experiments also did not generate better results than the second set of experiments (8949).

3.5 Experiments with the Morphosyntactic and Semantic Database Partitioned by Categories

Finally, the last set of experiments was performed with the base with morphosyntactic and semantic characteristics partitioned by categories. Table 10 presents the neural network confusion matrix, dictionary - first translation, dictionary - random translation and dictionary – in list translations, for the best result that occurred with 2 hidden layers and a learning rate of 0.01.

Table 10 - Matrix of confusion of the neural network and of the dictionaries in the morphosyntactic and semantic base partitioned by the categories.

Neural Network			Dictionary - First			Dictionary - Random			Dictionary - In List		
	0	1		0	1		0	1		0	1
0 (Real)	**5993**	34	0	**6039**	1	0	6038	2	0	6035	5
1 (Real)	166	**2938**	1	2356	664	1	2364	656	1	2267	**753**

Analyzing only the negative examples of the neural network (5993), it is possible to verify that the amount of correctness continued with a slightly inferior result compared with dictionaries. In addition, there was no gain compared to the best result for negative examples, found in the first set of experiments (6024).

Observing the positive examples of the neural network (2938), it is possible to say that the amount of correctness also continued much higher than the use of dictionaries. However, these experiments did not outperform the best result, for positive examples, found in the second set of experiments (2940).

By adding up the number of correct positive and negative examples (8931), it is verified that this set of experiments also did not generate better results than the second set of experiments (8949).

4. Conclusion

Interesting results were obtained through the idea of using a neural network, as a machine learning model, to help in the process of discovering equivalence of concepts learned by NELL in different languages.

Results generated by the experiments of the morphosyntactic database were almost 4 times better than the simple use of dictionaries. This result was fundamental to verify the viability of the research. Only after obtaining good results with the morphosyntactic basis has it been thought to expand the quantity of attributes.

We considered using the base with semantic attributes with the expectation that the insertion of this information, specific to each domain, will help the neural network in the process of discovering characteristics relevant to find equivalence of concepts.

For the time being, it was not possible to use only the semantic information to assist in the content equivalence discovery process. Some of the factors that contribute to this are related to: lower amount of data in Portuguese KB and dependence on the use of the dictionary for translation.

It is possible to cite at least two factors that contribute to the smaller size of the Portuguese base. The first one, is related to the amount of content. There are many more pages written in English, for NELL to use as a source of learning, than in Portuguese. The second, is due to the fact that the English base was created before the Portuguese base. The learning mechanisms of NELL do not occur instantaneously. In this way, having more running time favors having more information on the KB.

Another factor that negatively influences the result obtained in the semantic base, is related to the dependence of the translation of the values of the relations, exist in the dictionary. When this does not happen, the term is discarded.

For now, in addition to the use of semantic data along with the morphosyntactic data does not bring gain and it did bring some noise into the neural network. In this way, future work involves the use of larger sets of existing information, based on information in Portuguese and in English, that NELL does not yet consider as fact, but which has a high level of confidence that it is correct. With more information it is expected that the number of attributes with zeros will decrease and the neural network will achieve better results in the process of finding characteristics relevant to classification. After that, it will be possible to make a detailed analysis of the results found and make comparisons with other models found in the literature.

5. References

Duarte M.C. (2011). "Aprendizado Semissupervisionado através de técnicas de acoplamento", https://repositorio.ufscar.br/bitstream/handle/ufscar/474/3777.pdf.

Duarte M.C. (2014). Exploring two Views of Coreference Resolution in a Never-Ending Learning System. In *International Conference on Hybrid Intelligent Systems (HIS)*.

González J., Hruschka E.R., Mitchell T.M (2017) "Merging Knowledge bases in different languages", http://www.aclweb.org/anthology/W17- 2403

Haykin, S. (1999) "Neural Networks: A Comprehensive Foundation", bookman, ed. 2, Hamilton, Ontario, Canada.

Ke Y., Hagiwara M. (2015). A Natural Language Processing Neural Network Comprehending English. In *International Joint Conference on Neural Networks (IJCNN)*, https://ieeexplore.ieee.org/abstract/document/7280492/

Kurzweil, R. (1990) "The Age of Spiritual Machines", The MIT Press, Massachusetts.

Mitchell, T.M. (1997) "Machine Learning", McGraw-Hill, 1. ed., New York, NY, USA.

Mitchell, T.M. et al (2018). Never-ending learning. In *Communications of the ACM*, v. 61, pages 103-115.

Wijaya, D.T., Mitchell T.M. (2016) Mapping Verbs in Different Languages to Knowledge Base Relations using Web Text as Interlingua. In: HLT-NAACL.

Zhu, X. et al (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Machine Learning-International Workshop Then Conference*, v. 20, n. 2, page 912.