Generic Ontology Learners on Application Domains

Francesca Fallucchi, Maria Teresa Pazienza, Fabio Massimo Zanzotto

DISP - University of Rome “Tor Vergata” (Italy)
{fallucchi,pazienza,zanzotto}@info.uniroma2.it

Abstract

In ontology learning from texts, we have *ontology-rich* domains where we have large structured domain knowledge repositories or we have large general corpora with large general structured knowledge repositories such as WordNet (Miller, 1995). Ontology learning methods are more useful in *ontology-poor* domains. Yet, in these conditions, these methods have not a particularly high performance as training material is not sufficient. In this paper we present an LSP ontology learning method that can exploit models learned from a generic domain to extract new information in a specific domain. In our model, we firstly learn a model from training data and then we use the learned model to discover knowledge in a specific domain. We tested our model adaptation strategy using a *background* domain that is applied to learn the *isa* networks in the Earth Observation Domain as a specific domain. We will demonstrate that our method captures domain knowledge better than other generic models: our model better captures what is expected by domain experts than a baseline method based only on WordNet. This latter is better correlated with non-domain annotators asked to produce the ontology for the specific domain.

1. Introduction

Domain knowledge bases are extremely important in a variety of natural language processing applications but manually creating structured knowledge repositories is a very time consuming and expensive task. Semi-supervised learning of domain knowledge bases from texts is generally seen as the solution. This is a very attractive and rich research area that is full of challenges. Generally, the process for automatically creating, adapting, or extending existing knowledge bases relies on existing structured knowledge and domain corpora. In ontology learning models using lexico-syntactic patterns (LSP) (Robison, 1970; Hearst, 1992a; Pantel and Pennacchiotti, 2006), existing domain ontologies or structured knowledge bases give positive learning examples. These latter are exploited to learn lexico-syntactic patterns from domain corpora. Learnt LSPs are then used to extract and structure new knowledge from the domain corpora. For a successful application, these LSP methods for learning domain ontologies need large domain corpora and existing domain knowledge bases. LSP methods for learning ontologies from texts are good models only when we consider *ontology-rich* domains or we do generic knowledge extraction. In this latter case, these methods can exploit large general corpora and large general structured knowledge repositories such as WordNet (Miller, 1995). There are only few domains with well-assessed existing structured knowledge bases where the problem is to expand these ontologies. On the contrary, the large number of applications domains has little or not existing structured knowledge. The big challenge is to successful apply these methods in *ontology-poor* domains.

One of the possible ways to address the above challenge is to build LSP models that learn lexico-syntactic patterns on generic and ontology rich domains and then apply these patterns on specific ontology poor domains. In line with (Gao et al., 2009), we respectively refer as the *background domains* and *application domains* to these two kinds of domains. Yet, in machine learning and in statistical learning learning data should be enough representative of the environment where learned models will be applied. The statistical distribution of learning data should be similar to the distribution of the data where the learn model is applied. In this application scenario, this assumption is inaccurate. *Background domain data*, also called out-of-domain data, used for learning lexico-syntactic patterns have generally a different distribution with respect to *application domain data*, also called in-domain data. Generally, out-of-domain data are more than in-domain data. We need to envisage methods that exploit these data for building accurate in-domain models.

In this paper we present an LSP ontology learning method that can exploit models learned from a generic domain to extract new information in a specific domain. In our model, we firstly learn a model from training data and then we use the learned model to discover knowledge in a specific domain. In line with (Gao et al., 2009), we call *background* domain that we use for training purposes. The background domain is a generic domain defined through a generic corpus and a generic knowledge base. The *adaptation* domain is the domain where we apply the model. The *adaptation* domain corpus is used to generate feature vectors for each domain pair. The learned model will decide if the a ontological relation between two words hold in the particular domain. We tested our model adaptation strategy using a *background* domain that is applied to learn the *isa* networks in the Earth Observation Domain as a specific domain. We will demonstrate that our method captures domain knowledge better than other generic models: our model better captures what is expected by domain experts than a baseline method based only on WordNet. This latter is better correlated with non-domain annotators asked to produce the ontology for the specific domain.

The rest of the paper is organized as follows. In section 2., we analyze the related work on the area of domain adaptation. Then, we present our model in Section 3.. In section 4., we, then, evaluate and assess the performance of our method on the target domain, i.e., Earth Observation Domain. Finally, in section 5., we draw some conclusions.

2. Related Work

One of the basic assumptions in machine learning and in statistical learning is that learning data are enough representative of the environment where learned models will be applied. The statistical distribution of learning data should be similar to the distribution of the data where the learn model is applied. In natural language processing tasks involving semantics, this assumption is extremely important. Learning ontologies from texts using lexico-syntactic pattern (LSP) based methods is one of these semantic tasks. LSP methods (Hearst, 1992b; Pantel and Pennacchiotti, 2006; Snow et al., 2006) generally use existing ontological resources to extract learning examples. These latter are matched over a collections of documents to derive lexico-syntactic patterns describing a semantic relation. These patterns are then used to expand the existing ontological resource by retrieving and selecting new examples.

LSP ontology learning methods are generally used to expand existing domain ontologies using domain corpora or to expand generic lexical resources (e.g., Wordnet (Miller, 1995)) using general corpora (Snow et al., 2006; Fallucchi and Zanzotto, 2009b). In this way, the basic assumption of machine learning approaches is satisfied. Yet, the nature of the ontology learning task requires that models learned in a general or a specific domain may be applied in other domains for building or expanding poor initial ontologies using domain corpora. In this case, the distribution of learning and application data is different. Learned LSP models are “domain-specific” as lexico-syntactic patterns may be related to the prose of a specific domain. These models are then accurate for the specific domain but may fail in other domains. If the target domain has not relevant pre-existing ontologies to expand, we will not have enough data for training the initial model. In (Snow et al., 2006), all WordNet has been used as source of training examples. In these cases, we need to adopt domain adaptation techniques (Gildea, 2001; Roark and Bacchiani, 2003; Chelba and Acero, 2006; Daumé and Marcu, 2006; Gao et al., 2009; Bacchiani et al., 2004).

Domain adaptation is a known problem in machine learning and statistical learning. To stress the difference between the distribution of the data in the original domain (also called *background domain*) and in the target domain, these two are referred as *out-of-domain* data and as *in-domain* data. *Out-of-domain* data are generally large sets and are used for training.

In the general application scenario for ontology learning method the assumption that out-of-domain data and in-domain data share the same underlying probability distribution is inaccurate. This happens in many applications. Generally, in-domain data is drawn from a distribution that is related, but not identical, to out-of-domain distribution of the training data. As out-of-domain data are generally more than in-domain data, we need to envisage methods that exploit these data for building accurate in-domain models.

The domain adaptation problem exactly consists of leveraging out-of-domain data to derive models performing well on in-domain data.

This is a natural need as manually building initial training resources for new domains is an expensive task just as designing a system for each target domain. Then the natural expectation is to minimize the amount of effort required to building in-domain data using a model trained with out-of-domain data. In such cases, it becomes very important to adapt existing models from rich source domains to resource-poor target domains.

The problem of domain adaptation arises in a variety of applications in natural language processing (Blitzer et al., 2006; Chelba and Acero, 2006; Daumé and Marcu, 2006): machine translation (Bertoldi and Federico, 2009), word sense disambiguation (Chan and Ng, 2007) and many other areas.

Different domain adaptation techniques are introduced in the context of specific applications and statistical learning methods. For example, a standard technique used in statistical language modeling and in other generative models is the maximum a posteriori (MAP) estimation (Gauvain and Lee, 1994) as prior knowledge to estimate the model parameters. The MAP framework is general enough to include some previous model adaptation approaches, such as corpus mixing (Gildea, 2001). Another example is the MAP estimation used in (Roark and Bacchiani, 2003) to adapt a lexicalized probabilistic context-free grammar (PCFG) to a novel domain. In (Chelba and Acero, 2006) a MAP adaptation technique for maximum entropy models is developed for the problem of recovering the correct capitalization of uniformly case text for language modeling in speech recognition. In (Daumé and Marcu, 2006) a statistical formulation is provided, that is a mixture of maximum entropy model and linear Chain Models for conditional random fields. Two other classes of model adaptation methods are very interesting: error-driven learning approaches and model interpolation approaches. In error-driven learning approaches, the background model is adjusted to minimize the ranking errors made by the model on the adaptation data (Gao et al., 2009; Bacchiani et al., 2004). In model interpolation approaches, the in-domain data are used to derive an adaptation model, which is then combined with the background model trained on out-of-domain data. In (Gao et al., 2009) model interpolation is investigated for web search ranking.

One of the possible ways of using the model adaptation is to adjust the model trained on the background domain to a different domain (the adaptation domain) modifying opportunely the parameters and/or structure. The motivation of this approach is that usually the background domain has large amounts of training data while the adaptation domain has only small amounts of data.

In this paper we propose an approach to the domain adaptation problem where we build a background model and we use its predictions as features for the in-domain data. The basic idea is that in-domain data can be obtained to adapt all components of an already developed system.

In (Gao et al., 2009) a set of error-driven learning methods are developed where, in an incremental way, each feature weight could be changed separately but also new features could be constructed.

In our case, differently, we do not add any feature but we only change each feature weight in accord to the used in-domain corpus.

In (Blitzer et al., 2006) a common representation for features extracted from different domains is given using pivot features from unlabeled data to put domain-specific words in correspondence. Pivot features are features which occur frequently in the two domains and behave similarly in both. By analogy with (Blitzer et al., 2006) we propose to learn common features, meaningful for both domains that have different weights in accord to the occurrences in the different corpus used for the two domains. We hypothesize that a model trained in the source domain using this common feature representation will generalize better the target domain. In some cases, many steps may be required to adapt a model trained on the source domain for use in the target domain (Daumé and Marcu, 2006; Roark and Bacchiani, 2003; Ando, 2004). On the contrary, in our work we learn a model from the out-of-domain data and we use it to learn the in-domain data without any additional effort.

3. Learner Model: from Background to Application domain

Can training data from one corpus be applied to learn another corpus? The basic idea is partly to answer this question because we want to define an ontology learning model that can be adapted to previously unseen distributions of data. This model is thought to exploit the information learned in a *background* domain for extracting information in an *adaptation* domain.

Our ontology learning method is based on the probabilistic formulation given in (Snow et al., 2006; Fallucchi and Zanzotto, 2009a). We use this probabilistic setting to learn a model that takes into consideration corpus-extracted evidences over a list of training pairs. The initial feature space is built starting from the analysis of a generic corpus where we observe a list of training pairs of words that are in a target semantic relation. We can generate these pairs using general resources such as WordNet. These pairs are used to enable the probabilistic method to induce lexico-syntactic patterns for the model of the specific semantic relation (Hearst, 1992b). The learned model can be used to estimate the probabilities of the new instances computing a new feature space using the corpus of the *adaptation* domain.

In the rest of this section, we will firstly describe the background ontology learning model (Sec. 3.1.) and we will then illustrate the method that we will be adapted to the new domain (Sec. 3.2.).

3.1. Background Ontology Learner

In the probabilistic formulation, the task of learning ontologies from a corpus is seen as a maximum likelihood problem. The ontology is seen as a set O of assertions R over pairs $R_{i,j}$. In particular we will consider the *is-a* relation. In this case, if $R_{i,j}$ is in O, i is a concept and j is one of its generalizations. For example, $R_{dog,animal} \in O$ states that dog is an $animal$ according to the ontology O.

The main probabilities are then: (1) the prior probability $P(R_{i,j} \in O)$ of an assertion $R_{i,j}$ to belong to the ontology O and (2) the posterior probability $P(R_{i,j} \in O|\overrightarrow{e}_{i,j})$ of an assertion $R_{i,j}$ to belong to the ontology O given a set of evidences $\overrightarrow{e}_{i,j}$ derived from the corpus. These evidences are derived from the contexts where the pair (i, j) is found in the corpus. The vector $\overrightarrow{e}_{i,j}$ is a feature vector associated to a pair (i, j). For example, a feature may describe how many times i and j are seen in patterns like *"i as j"* or *"i is a j"*. But many other indicators exist of an Is-a relation between i and j (see (Hearst, 1992b)).

Given a set of evidences E over all the relevant word pairs, the probabilistic ontology learning task is defined as the problem of finding an ontology $\widehat{O}$ that maximizes the probability of having the evidences of E, i.e.:

$$\widehat{O} = \arg\max_{O} P(E|O)$$

In the original model (Snow et al., 2006; Fallucchi and Zanzotto, 2009a), this maximization problem was solved by a local search.

In the present model at each step we maximize the ratio between the likelihood $P(E|O')$ and the likelihood $P(E|O)$ where $O' = O \cup N$ and N are the relations added at each step. As in (Snow et al., 2006; Fallucchi and Zanzotto, 2009a) this ratio is called *odds*. It is calculated using the logistic regression and then solving a linear problem using the pseudo-inverse matrix (Fallucchi and Zanzotto, 2009a). The regression coefficients will be estimated as follows

$$\widehat{\beta} = X^{+}_{C_B} l \tag{1}$$

where l is the logit vector and $X^{+}_{C_B}$ is the **Moore-Penrose pseudoinverse** (Penrose, 1955) matrix of the inverse evidence matrix X_{C_B} obtained from a generic corpus C_B that includes a constant column of 1's, necessary to obtain the β_0 coefficients.

The regressors represent the model that we learned from the training pairs using a generic corpus C_B that we will use to compute the probabilities of the testing pairs.

3.2. Estimator for Application Domain

In our task, instead of finding the ontology that maximizes the likelihood of having the evidences E, we calculate, given the regressors, the probabilities of the testing pairs step by step. The idea is that, given the domain based corpus C_A, for each testing pair we compute the vector space according to the features selected in the previous generic corpus feature space analysis.

After the domain based corpus feature space analysis where we look for the testing pairs in C_A, we obtain a new feature space X_{C_A}. It is a matrix $n' \times m$ where n' is the number of the new instances found in the corpus C_A and m is the number of the features. We compute the logit of the new instances as in (Fallucchi and Zanzotto, 2009a)

$$l' = \alpha X_{C_A} \widehat{\beta} \tag{2}$$

where X_{C_A} is the inverse evidence matrix obtained from a *adaptation* domain corpus C_A that includes a constant column of 1's, necessary to obtain the β_0 coefficients. The parameter α is used to adapt the model by the β vector to the new domain.

From the definition of logit we can compute the probabilities of the new instances, i.e.:

$$p_i = \frac{\exp(l_i)}{1 + \exp(l_i)} \tag{3}$$

This latter can be used to build the know ledge base in the new domain.

4. Experimental Evaluation

We experimented with our model adaptation strategy using a generic domain as *background* domain and the Earth Observation Domain as specific domain. We took the isa relation as the target relation. The target of the experiments is to understand whether or not our model adapt to specific domains. We then compare our system (Our-System) with respect to a system that uses only WordNet (WN-System).
In this section, we firstly describe the general experimental set up. We then describe the quality of the target domain ontologies. Finally, we analyze the accuracy of our models with respect to the three different ontologies.

4.1. Experimental Setup

To define completely the experiments we have to define: both training and testing pairs, which corpus has been used to extract evidences for training pairs, which corpus to extract evidences for testing pairs, and which feature space we use for both corpora.
To build the training pairs we generated all the pairs that were in hyperonym relation in WordNet[1] (Miller, 1995) and we obtained about 2 millions of words.
Here, we firstly define the semantic networks used in the experiments of Section 4.3.. The network of words will be used as a source of training and testing examples. For each experiment we need: a training example set $TR = (TR_p, TR_n)$ with positive pairs TR_p and negative pairs TR_n, and a testing example set TS.
To build TS we start from a given list of 63 terms that are relevant in Earth Observation Domain. Then we combine each term with the other terms and we generate 63 × 63 pairs. Furthermore, for each term w, we select all the synsets s_w in WordNet. In the case of a term with a synset in WordNet we generate the pairs combining w with all the hyperonyms for each synset. Otherwise, if w has compound words we look for our semantic head in WordNet. If we find the synsets, we generate the pairs combining w with the hyperonyms of the semantic head of w.
We extract the training example pairs from an existing knowledge repository: WordNet[2] (Miller, 1995).
Given hyperonymy as target relation, we can derive the network of words $\mathcal{R}$ from the set R as follows: $\mathcal{R} = \{(w_a, w_b)|(S_a, S_b) \in R, w_a \in S_a, w_b \in S_b\}$. We then build the set $\mathcal{H}$ that contains all pairs of words in WordNet that are in hyperonymy relation. Then $TR_p = \mathcal{H} - \mathcal{TS}$.
Given the set of the words in WordNet W, the training negative example is $TR_n = W \times W - TR_p - TS$. We build TR_p, TR_n and TS without overlap.
We searched for the pairs in TR in a corpus C_B (in particular the *English Web as Corpus* (ukWaC) (Ferraresi et al., 2008) has been used). This is a web extracted corpus of about 2700000 web pages containing more than 2 billion words. It contains documents of several different topics such as web, computers, education, public sphere, etc.. It has been largely demonstrated that the web documents are good models for natural language (Lapata and Keller, 2004).
Using a web crawler, here we pick up a corpus related to Earth Observation Domain C_A , successively "cleaned", that contains about 8300 documents (115,6 MB).
We use the bag-of-word feature space. Out of the $T \cup \overline{T}$, only those pairs that appeared at a distance of 3 tokens at most have been selected. Using these 3 tokens, we generate the *bag-of-word* feature space.
The pairs in TR found in the ukWaC are 527348, while the pairs in TS found in C_A are 404. The two generated feature spaces have the same features that are 276670.
The model to build ontologies in Earth Observation Domain has been generated by using the training pairs and the corpus ukWac.

4.2. Evaluating the Quality of Target Domain Specific Ontologies

We want to evaluate our approach in learning the bulk of the ontologies, i.e., the *isa* relation, in Earth Observation Domain. between two pairs of words is a binary problem. We then asked three annotators (A_1, A_2 and A_3) to build three different ontologies: two of them are expert in the domain (A_1 and A_2), the third one is not (A_3). A_1 and A_2 have different levels of expertise: A_1 is a young expert in the domain and A_2 an older one. Each annotator made a binary classification of 641 pairs of words in Earth Observation Domain, i.e., the TS set introduced in the previous section.
We then wanted to judge the quality of the annotation procedure according to their inter annotation agreement. A simple measure of the quality of the agreement rate between two human annotators is the ratio between the number of items identically judged by two different annotators and the total number of items considered by the annotators. In (Scott, 1955) this measure is named **observed agreement** A_o and it is defined as *the percentage of judgments on which the two analysts agree when coding the same data independently*. In accord to (Artstein and Poesio, 2008) we define the agreement value agr_i for all items I as follows:

$$arg_i = \begin{cases} 1 & \text{if annotators assign i to the same category} \\ 0 & \text{if annotators assign i to different categories} \end{cases}$$

The observed agreement has been evaluated as in the following:

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i$$

This measure does not take into account changes in the agreement between two annotators. An improved measure of inter-annotator agreement is given by the Cohen's **kappa** coefficient (Cohen, 1960). It is a statistical measure that takes into account the effect of changes in the agreement giving the possible agreement beyond change actually observed. The kappa-coefficient is defined as follows:

$$k = \frac{A_o - A_e}{1 - A_e}$$

- A_e : expected agreement by change

[1] We used the version 3.0 of WordNet

[2] We use the version 3.0 in prolog.

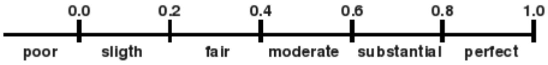

Figure 1: Scale for the interpretation of Kappa by Landis and Koch (1977)

- $1 - A_e$: attainabled agreement over and above change
- $A_o - A_e$: actually found agreement beyond change

The expected agreement (A_e) is the probability of the agreement among annotators due to change. There are two different methods for estimating a probability distribution for random assignment of categories. The two approaches reflect different conceptualizations of the problem.

In the first approach, each annotator has a personal distribution, based on that annotator's distribution of categories (Cohen, 1960). In the second approach, there is one distribution for all annotators, derived from the total proportions of categories assigned by all annotators (Scott, 1955; Fleiss and others, 1971). Data are respectively visualized in a *contingency table* (first approach) and in an *agreement table* (second approach).

The distinction between the two approaches, in the case of two annotators, is often glossed over because in practice the two computations of A_e produce very similar (when not the same) outcomes, as shown in section 4.2.1..

In (Carletta, 1996) the adaptation of the kappa coefficient to computational linguistic is suggested.

Different levels of agreement may be defined, according to the experiments of a specific application. In (Landis and Koch, 1977) confidence intervals are proposed for the values of the kappa coefficient, as reported in Figure 1.

We can examine the issue of inter-annotator agreement by comparing the agreement rate of the human annotators. There are different methods for measuring the agreement among 3 annotators.

When there are more than two annotators, some of them may agree and the rest disagrees on the same item. In this case, the observed agreement can no longer be defined as the percentage of items getting agreement. To solve this problem , we can analyze two solutions : **pairwise agreement** and **multi-π agreement** both in (Fleiss and others, 1971).

In the first section 4.2.1. we will describe the inter-annotators agreement for each pair of annotators that has a personal distribution and we will show that this is similar to the distribution computed on both annotators of each pair.

In the multi-π agreement, we examine the distribution of all the three annotators.

4.2.1. Pairwise agreement

The pairwise agreement defines the agreement on a particular item as the proportion of agreed judgment pairs out of the total number of judgment pairs for that item (Fleiss and others, 1971).

We measure the inter-annotators agreement of the 3 pairs of annotators: $pair_1$ for the two annotators expert in the domain A_1 and A_2; $pair_2$ for one annotator expert in the domain A_1 and the other one not expert A_3; and, $pair_3$ for the second annotator expert in the domain A_2 and the other one not expert A_3.

Each annotator annotates 641 pairs of words in Earth Observation Domain and assigns to each pair one of the two labels "YES" or "N0" (1 or 0). It is important to note that only 404 pairs are found in Earth Observation Domain corpus. We name 641-annotations the list that contains 641 annotations of each annotator and 404-annotations the list that contains 404 annotations of each annotator. In the following we discuss the inter-annotator agreement with respect to both the lists of annotations.

Given the same data (641 or 404-annotations) with the same guidelines, we build the contingency tables for the 3 pairwise annotators(respectly Table 2 and Table 4). For each table we report the statistic of the two annotators.

Then in Table 1a we summarize the inter-annotator agreement of the 3 pairwise agreements considering 641-annotators. For example, the observed agreement for this data is obtained summing up the cells of the table where the annotators assign the same judgement and dividing by the total number of annotations.

For example, considering $pair_1$ (first row of the Table 1a), the two annotators label 47 occurrences as YES, and 490 as NO. The resulting observed agreement of $pair_1$ is $A_o = (47 + 490)/641 = 0.8377535$. As above mentioned, there are two different methods to compute the expected agreement. In the first method the expected agreement is governed by prior distributions that are unique for each annotator and it is computed looking the actual distribution. Then for $pair_1$ we have $A_e = 0.16848674 * 0.1404056 + 0.83151326 * 0.8595944 = 0.7384206$.

In the second method we get the same distribution for each annotator of the $pair$, then we have

$$A_e = \left(\frac{90 + 108}{641 * 2}\right)^2 + \left(\frac{533 + 551}{641 * 2}\right)^2 = 0.7388149$$

Since the two A_e values are similar and the same occurs for the other pairs, we report only the expected agreement computed using the first method

Finally, using both the observed and expected agreement, the possible agreement beyond change observed for the $pair_1$ is $kappa = (0.8377535 - 0.7384206)/(1 - 0.7384206) = 0.3797428$. Analogously we compute kappa value for the other pair of annotators.

In the same way we compute Observed Agreement, Expected Agreement and coefficient kappa for the pairwise agreement considering 404-annotations (Table 3a). Summarizing only for $pair_3$ on 641-annotations the coefficient kappa is in the "fair" interval in accord to the scale proposed in (Landis and Koch, 1977) and reported in Figure 1. Most likely there is a fair agreement between annotators A_2 and A_3 because the first one is an older expert in the domain while the second one is not expert at all, so they have a different knowledge with respect to the specific Earth Observation Domain.

In all the other cases the pairwise agreement is better because the coefficient kappa belongs to the "moderate" interval. We are confident on the reliability of such annotations

		A_1		
		yes	no	
A_2	yes	47	61	108
	no	43	490	533
		90	551	641

(a) $pair_1 = (A_1, A_2)$

		A_1		
		yes	no	
A_3	yes	76	83	159
	no	14	468	482
		90	551	641

(b) $pair_2 = (A_1, A_3)$

		A_2		
		yes	no	
A_3	yes	72	87	159
	no	36	446	482
		180	533	641

(c) $pair_3 = (A_2, A_3)$

Table 1: Contingency tables for pairwise annotator agreement for 641-annotations

	A_o	A_e	$kappa$
$pair_1 = (A_1, A_2)$	0.8377535	0.7384206	0.3797428
$pair_2 = (A_1, A_3)$	0.8486739	0.6811997	0.5253266
$pair_3 = (A_2, A_3)$	0.8081123	0.6670496	0.4236749

Table 2: pairwise agreement for 641-annotations

as the annotators agree on labeling the same pairs of words. This allows us to prove the validity of the annotation.

4.2.2. Multi-π agreement

In multi-π agreement the agreement of the annotators is considered as a whole. There is only one distribution for all the annotators, derived from the total proportions of categories assigned by each annotator.

When there are more than two annotators, the visualization of the data is a difficult task: a possible solution is in using the agreement table where each annotator is represented in a separate column.

The columns A_1, A_2, and A_3 of table 4a and table 4b report the label 1 or 0 assigned for each pair (first column) by the 3 annotators respectively in 641 or 404-annotations.

For both tables we report in the columns YES and NO respectively the sum of 1s and 0s in A_1, A_2, and A_3. In table 4c we report the observed and expected agreement and the relative kappa coefficient for both 641 and 404 annotations. The kappa value obtained from both annotations confirms the conclusions deduced with the pairwise agreement method that proved the validity of the annotations of the 3 annotators.

4.3. Result

In our experiments we investigated how the approach to compute a model using both a *background* domain and an existing network, can be positively used to learn the *isa* relation in Earth Observation Domain.

For the evaluation, we compare our learner model (*Our-System*) directly with currently existing hyperonym links in WordNet (*WN-System*) and we measure in both cases the performance to find correctly the testing pairs that are in isa relation.

In order to evaluate the performance of the two systems for the pairs in Earth Observation Domain we used the three different ontologies produced by the three annotators. We will call these three target ontologies with the name of the annotator.

The results of the experiments are reported in Table 6a and in Table 6b. In the first table we compute the recall, the precision and the f-measure of the *WN-System* against the 3 ontologies, while in the second table we compute the recall, the precision and the f-measure of the *Our-System*.

We can then draw some observations: First, *Our-System* behaves better than the *WN-System* on the ontologies produced by the expert annotators. The f-measure of both the expert annotators (A_1 and A_2) is better for *Our-System* with respect to *WN-System*. On the contrary, for the last ontology (A_3) the *WN-System* has better performance than our system. Then, our system is capturing knowledge of the specific domain as it is behaving better than the generic system with respect to domain experts. Second, in the case of the expert annotators, the recall of our system is higher than the recall of the WordNet based system. This confirms that the coverage of WordNet in the specific domain is low and only learning methods can be used to adapt the ontological information to the specific domain. On the contrary, for the non-domain expert, WordNet is good enough to cover domain knowledge.

In conclusion, results show that *Our-System* is a good learner method that can be positively used to learn the *isa* relation in Earth Observation Domain.

5. Conclusion

In this paper we present an ontology learning method that can exploit the models learned from a generic domain to extract new information in a specific domain. In our model, we firstly learn a model from the training data, then we use the learned model to discover the relation between two words in a specific domain.

We tested our model adaptation strategy using a *background* domain that is applied to learn the *isa* networks in a specific domain, i.e., the Earth Observation Domain. The results of the experiments are promising showing that this way of using a model identified in a *background* domain is helpful to learn the *isa* relation in Earth Observation Domain.

		A_1		
		yes	no	
A_2	yes	40	32	72
	no	35	297	332
		75	329	404

(a) $pair_1 = (A_1, A_2)$

		A_1		
		yes	no	
A_3	yes	65	54	119
	no	10	275	285
		75	329	404

(b) $pair_2 = (A_1, A_3)$

		A_2		
		yes	no	
A_3	yes	53	66	119
	no	19	266	285
		72	332	404

(c) $pair_3 = (A_2, A_3)$

Table 3: Contingency tables for pairwise annotator agreement for 404-annotations

	A_o	A_e	$kappa$
$pair_1 = (A_1, A_2)$	0.8341584	0.7023086	0.4429077
$pair_2 = (A_1, A_3)$	0.8415842	0.6291663	0.5728117
$pair_3 = (A_2, A_3)$	0.7896040	0.6322174	0.4279336

Table 4: pairwise agreement for 404-annotations

pairs of words	A_1	A_2	A_3	Yes	NO
(agriculture,department)	0	0	0	0	3
(soil,earth)	1	1	1	3	0
(agriculture,business)	0	0	0	0	3
(wind,direction)	1	0	0	1	2
(climate,climate change)	0	0	0	0	3
(climate change,climate)	0	1	1	2	1
(climate change,activity)	1	0	1	2	1
(forest,terra firma)	1	1	1	3	0
…	…	…	…	…	…
TOTAL	90	108	159	357 (0.19)	1566 (0.81)

(a) Agreement table for 641-annotations

pairs of words	A_1	A_2	A_3	Yes	No
(forest,terra firma)	1	1	1	3	0
(wind,process)	0	0	0	0	3
(forest,object)	0	0	0	0	3
(cloud,state)	0	1	0	1	2
(soil,object)	0	1	1	2	1
(wind,breath)	0	0	0	0	3
(wind,act)	0	0	0	0	3
(topography,geography)	1	1	1	3	0
…	…	…	…	…	…
TOTAL	75	72	119	266 (0.22)	946 (0.78)

(b) Agreement table for 404-annotations

	A_o	A_e	$kappa$
641-annotations	0.83151	0.69764	0.44277
404-annotations	0.82382	0.65739	0.48577

(c) Multi-π agreement rispet to 641 and 404 annotations

Table 5: Agreement tables and Multi-π agreement for 641 and 404 annotations

annotators	recall	precision	f-measure
A_1	0,36	0.184932	0,244344
A_2	0,305556	0,150685	0,201836
A_3	0,470588	0,383562	0,422642

(a) *WN-System* against the 3 annotators

annotators	recall	precision	f-measure
A_1	0,493333	0,253425	0,334842
A_2	0,4305556	0,212329	0,284404
A_3	0,4369748	0,356164	0,392453

(b) *Our-System* against the 3 annotators

Table 6: Performance of both systems with respect to 3 annotators

Annotators	recall	precision	f-measure
A_1	0,36	0.184932	0,244344
A_2	0,305556	0,150685	0,201836
A_3	0,470588	0,383562	0,422642

(a) *WN-System* against the 3 annotators

Annotators	recall	precision	f-measure
A_1	0,493333	0,253425	0,334842
A_2	0,4305556	0,212329	0,284404
A_3	0,4369748	0,356164	0,392453

(b) *Our-System* against the 3 annotators

Table 7: Performance of both systems with respect to 3 annotators

6. References

Rie Kubota Ando. 2004. Exploiting unannotated corpora for tagging and chunking. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, Morristown, NJ, USA. Association for Computational Linguistics.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Journal of Computational Linguistics*, 34(4).

Michiel Bacchiani, Brian Roark, and Murat Saraclar. 2004. Language model adaptation with map estimation and the perceptron algorithm. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 21–24, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 182–189, Morristown, NJ, USA. Association for Computational Linguistics.

John Blitzer, Ryan Mcdonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56, Prague, Czech Republic, June. Association for Computational Linguistics.

Ciprian Chelba and Alex Acero. 2006. Adaptation of maximum entropy capitalizer: Little data can help a lot.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Psychological Bulletin*, 20:37–46.

Hal Daumé, III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, May.

Francesca Fallucchi and Fabio Massimo Zanzotto. 2009a. SVD feature selection for probabilistic taxonomy learning. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 66–73, Athens, Greece, March. Association for Computational Linguistics.

Francesca Fallucchi and Fabio Massimo Zanzotto. 2009b. Svd for feature selection in taxonomy learning. In *Proceedings of the Conference on Recent Advances on Natural Language Processing*. John Benjamins, 14-16 September.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the WAC4 Workshop at LREC 2008*, Marrakesh, Morocco.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jianfeng Gao, Qiang Wu, Chris Burges, Krysta Svore, Yi Su, Nazan Khan, Shalin Shah, and Hongyan Zhou. 2009. Model adaptation via model interpolation and boosting for web search ranking. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 505–513, Morristown, NJ, USA. Association for Computational Linguistics.

Jean-luc Gauvain and Chin-hui Lee. 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2:291–298.

Daniel Gildea. 2001. Corpus variation and parser performance.

Marti A. Hearst. 1992a. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 15th CoLing*, Nantes, France.

Marti A. Hearst. 1992b. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics (CoLing-92)*, Nantes, France.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March.

Mirella Lapata and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of nlp tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, MA.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, November.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia, July. Association for Computational Linguistics.

R. Penrose. 1955. A generalized inverse for matrices. In *Proc. Cambridge Philosophical Society*.

Brian Roark and Michiel Bacchiani. 2003. Supervised and unsupervised pcfg adaptation to novel domains. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 126–133, Morristown, NJ, USA. Association for Computational Linguistics.

Harold R. Robison. 1970. Computer-detectable semantic structures. *Information Storage and Retrieval*, 6(3):273–288.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, 19(3):321–325.

Rion Snow, Daniel Jurafsky, and A. Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *In ACL*, pages 801–808.