Metrics for Measuring Domain Independence of Semantic Classes

Andrew Pargellis, Eric Fosler-Lussier, Alexandros Potamianos, Chin-Hui Lee

Dialogue Systems Research Dept., Bell Labs, Lucent Technologies Murray Hill, NJ, USA
{anp,fosler,potam,chl}@research.bell-labs.com

Abstract

The design of dialogue systems for a new domain requires semantic classes (concepts) to be identified and defined. This process could be made easier by importing relevant concepts from previously studied domains to the new one. We propose two methodologies, based on comparison of semantic classes across domains, for determining which concepts are domain-independent, and which are specific to the new task. The *concept-comparison* technique uses a context-dependent Kullback-Leibler distance measurement to compare all pairwise combinations of semantic classes, one from each domain. The *concept-projection* method uses a similar metric to project a single semantic class from one domain into the lexical environment of another. Initial results show that both methods are good indicators of the degree of domain independence for a wide range of concepts, manually generated for three different tasks: *Carmen* (children's game), *Movie* (information retrieval) and *Travel* (flight reservations).

1. Introduction

Despite the significant progress that has been made in the area of speech understanding for spoken dialogue systems, designing the understanding module for a new domain requires large amounts of development time and human expertise [5]. The design of speech understanding modules for a single domain (also referred to as a *task*) has been studied extensively in the literature [1, 2, 3, 4]. However, speech understanding models and algorithms designed for a single task, have little generalization power and are not portable across application domains. In this paper, the portability of semantics is investigated and measures of domain independence are proposed.

The first step in designing an understanding module for a new task is to identify the set of semantic classes, where each semantic class is a meaning representation, or *concept*, consisting of a set of words and phrases with similar semantic meaning. Some classes, such as those consisting of lists of names from a lexicon, are easy to specify. Others require a deeper understanding of language structure and the formal relationships (syntax) between words and phrases. A developer must supply this knowledge manually, or develop tools to automatically (or semi-automatically) extract these concepts from annotated corpora with the help of language models (LM). This can be difficult since it typically requires collecting thousands of annotated sentences, usually an arduous and time-consuming task.

One approach is to automatically extend to a new domain any relevant concepts from other, previously studied tasks. This requires a methodology that compares semantic classes across different domains. It has been demonstrated that semantic classes for a single domain can be semi-automatically extracted from training data using statistical processing techniques [6, 7, 8, 9, 10] because semantically similar phrases share similar syntactic environments [9, 10]. This raises an interesting question: Can semantically similar phrases be identified across domains? If so, it should be possible to use these semantic groups to extend speech-understanding systems from known domains to a new task. Semantic classes, developed for well-studied domains, could be used for a new domain with little modification.

We hypothesize that domain-independent semantic classes (concepts) should occur in similar syntactic (lexical) contexts *across* domains. We propose a methodology for rank ordering concepts by degree of domain independence. By identifying task-independent versus task-dependent concepts with this metric, a system developer can import data from other domains to fill out the set of task-independent phrases, while focusing efforts on completely specifying the task-dependent categories manually.

A longer-term goal for this metric is to build a descriptive picture of the similarities of different domains by determining which pairs of concepts are most closely related across domains. Such a hierarchical structure would enable one to merge phrase structures from semantically similar classes across domains, creating more comprehensive representations for particular concepts. More powerful language models could be built than those obtained using training data from a single domain.

2. Comparing concepts across domains

Semantic classes are typically constructed manually, using static lexicons to generate lists of related words and phrases. An automatic method of concept generation could be advantageous for new, poorly understood domains.[1] In this initial study, however, we validate our metrics using sets of predefined, manually generated classes.

We use two different statistical measurements to estimate the similarity of different domains. Figure 1 shows a schematic representation of the two metrics for a *movie* information domain (which encompasses semantic classes such as <CITY>, <THEATER NAME>, and <GENRE>), and a *travel* information domain (with concepts like <CITY>, <AIRLINE>, and <MONTH>).

The *concept-comparison* metric, shown at the top of Fig. 1, estimates the similarities for all possible pairs of semantic classes from two different domains. Each concept is evaluated in the lexical environment of its own domain. This method should help a designer identify which concepts are useful for many tasks, and which concepts could be merged into larger, more comprehensive classes.

[1] Several automatic or semi-automatic techniques exist for concept formation, such as building a context-free grammar (CFG) consisting of a set of phrase rules and semantic classes [11], or inducing them directly from statistical techniques [9, 10, 12, 13].

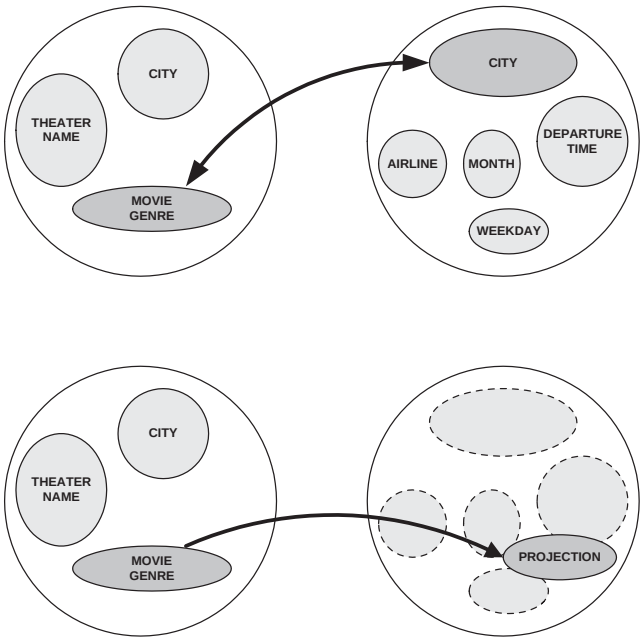

Figure 1: Pictorial view of the extension of the <GENRE> concept from the *Movie* domain (left) to the *Travel* domain (right). Top: **comparison** method. Bottom: **projection** method.

The *concept-projection* metric is quite similar mathematically to the concept-comparison metric, but it determines the degree of task (in)dependence for a single concept from *one* domain by comparing how that concept is used in the lexical environments of different domains. Therefore, this method should be useful for identifying the degree of domain-independence for a particular concept. Concepts that are specific to the new domain will not occur in similar syntactic contexts in other domains and will need to be fully specified when designing the speech understanding system.

2.1. Concept-comparison method

The comparison method compares how well a concept from one domain is matched by a second concept in another domain. For example, suppose (top of Fig. 1) we wish to compare the two concepts, <GENRE> = {*comedies*|*westerns*} from the *Movie* domain and <CITY> = {*san francisco*|*newark*} from the *Travel* domain. We do this by comparing how the phrases *san francisco* and *newark* are used in the *Travel* domain with how the phrases *comedies* and *westerns* are used in the *Movie* domain. In other words, how similarly are each of these phrases used in their respective tasks?

We develop a formal description by considering two different domains, d_a and d_b, containing M and N semantic classes (concepts) respectively.[2] The respective sets of concepts are $\{C_{a1}, C_{a2}, ..., C_{am}, ..., C_{aM}\}$ for domain d_a and $\{C_{b1}, C_{b2}, ..., C_{bn}, ..., C_{bN}\}$ for domain d_b. These concepts could have been generated either manually or by some automatic means. We find the similarity between all pairs of concepts across the two domains, resulting in $M \times N$ comparisons; two concepts are similar if their respective bigram contexts are similar. In other words, two concepts C_{am} and C_{bn} are compared by finding the distance between the contexts in which the concepts are found. The metric uses a left and right context bigram language model for concept C_{am} in domain d_a and the parallel bigram model for concept C_{bn} in domain d_b to form a probabilistic distance metric.

Since C_{am} is the label for the m^{th} concept in domain d_a, we use $\mathcal{W}_{am}$ to denote the set of all words or phrases that are grouped together as the m^{th} concept in domain d_a, i.e., all words and phrases that get mapped to concept C_{am}. As an example, C_{am} = <CITY> and $\mathcal{W}_{am}$ = {*san francisco* | *newark*}. Similarly, w_{am} denotes any element of the $\mathcal{W}_{am}$ set, i.e., $w_{am} \in \mathcal{W}_{am}$.

In order to calculate the cross-domain distance measure for a pair of concepts, we first replace in the training corpus d_a all instances of phrases $w_{am} \in \mathcal{W}_{am}$ with the label C_{am} (designated by $w_{am} \Rightarrow C_{am}$ for $m = 1..M$ in domain d_a and $w_{bn} \Rightarrow C_{bn}$ for $n = 1..N$ in domain d_b). Then a relative entropy measure, the Kullback-Leibler (KL) distance, is used to estimate the similarity between any two concepts (one from domain d_a and one from d_b). The KL distance is computed between the bigram context probability density functions for each concept. This KL distance is similar to the metric used in [9, 10], except that it uses domain-dependent probability distributions; the previous work cited only considers probability distribuitions within one domain.

We calculate the left and right language models, p^R and p^L; the left context-dependent bigram probabilities is of the form $p_a^L(v|C_{am})$, which can be read as "the probability that a word v is found to the *left* of any word in class C_{am} in domain d_a (*i.e.,* the ratio of counts of $\ldots vC_{am}\ldots$ to counts of $\ldots C_{am}\ldots$ in domain d_a). Similarly, the right context probability ($p_a^R(v|C_{am})$) is the probability that v occurs to the *right* of class C_{am} (equivalent to the traditional bigram grammar).

From these probability distributions, we can define KL distances by summing over the vocabulary V for a concept C_{am} from domain d_a and a concept C_{bn} from d_b. The *left* KL distance is given as

$$D^L_{am,bn} = D(p_a^L(C_{am}) \;||\; p_b^L(C_{bn})) = \\ = \sum_{v \in V} p_a^L(v|C_{am}) \;\log \frac{p_a^L(v|C_{am})}{p_b^L(v|C_{bn})} \quad (1)$$

and the right context-dependent KL distances are defined similarly.

The distance d between two concepts, C_{am} and C_{bn} is computed as the sum of the left and right context-dependent symmetric KL distances [10]. Specifically, the total symmetric distance between two concepts C_{am} and C_{bn} is

$$d(C_{am}, C_{bn}|d_a, d_b) = D^L_{am,bn} + D^L_{bn,am} + D^R_{am,bn} + D^R_{bn,am}$$

The distance between the two concepts C_{am} and C_{bn} is a measure of how similar their respective domains' lexical contexts are within which they are used [9, 10]. If our hypothesis is correct, similar concepts should have smaller KL distances. Larger distances indicate a poor match, possibly because one or both concepts are domain-specific. The comparison method enables us to compare two domains directly as it gives a measure of how many concepts, and which types, are represented in the two domains being compared. KL distances cannot be compared for different pairs of domains since they have different pair probability functions. So the absolute numbers are not meaningful, although the rank ordering within a pair of domains is.

[2] In the most general case, a "domain" could consist of concepts obtained for a merger of two or more previously studied domains.

2.2. Concept-projection method

The projection method investigates how well a single concept from one domain is represented in another domain. If the concept for a movie type is <GENRE> = {*comedies* | *westerns*}, we want to compare how the words *comedies* and *westerns* are used in both domains. In other words, how does the context, or usage, of each concept vary from one task to another? The projection method addresses this question by using the KL distance to estimate the degree of similarity for the same concept when used in the bigram contexts of two different domains.

As with the comparison method, the projection technique uses KL distance measures, but the distributions are calculated using the same concept for both domains. Since only a single semantic class is considered at a time for the projection method, the pdfs for both domains are calculated using the same set of words from just one concept, but using the respective LMs for the two domains. A semantic class C_{am} in domain d_a fulfills a similar function as in domain d_b if the bigram contexts of the phrases $w_{am} \in \mathcal{W}_{am}$ are similar for the two domains. In the projection formalism we replace words according to the two rules: $w_{am} \Rightarrow C_{am}$ for both the d_a and d_b domains. Therefore, both domains are parsed for the same set of words $w_{am} \in \mathcal{W}_{am}$ in the "projected" class, C_{am}. Following the procedure for the concept-comparison formalism, the left-context dependent KL distance $D^L_{am,bm}$ is defined as

$$D^L_{am,bm} = D(p^L_a(C_{am}) \,||\, p^L_b(C_{am})) = \\ = \sum_{v \in V} p^L_a(v|C_{am}) \, \log \frac{p^L_a(v|C_{am})}{p^L_b(v|C_{am})} \quad (2)$$

and the total symmetric distance

$$d(C_{am}, C_{am}|d_a, d_b) = D^L_{am,bm} + D^L_{bm,am} + D^R_{am,bm} + D^R_{bm,am}$$

measures the similarity of the same concept C_{am} in the different lexical environments of the two domains, d_a and d_b.

A small KL distance indicates a domain-independent concept that can be useful for many tasks, since the C_{am} concept exists in similar syntactical contexts for both domains. Larger distances indicate concepts that are probably domain-specific and do not occur in any context in the second domain. Therefore, projecting a concept across domains should be an effective measure of the similarity of the lexical realization for that concept in two different domains.

3. Preliminary results and discussion

In order to evaluate these metrics, we decided to compare manually constructed classes from a number of domains. We were hoping that the metrics would give us a rank-ordered list of the defined semantic classes, from task independent to task dependent. The evaluation was informal, relying on the experimenter's intuition of the task-dependence of the manually derived concepts.

Feature	Carmen	Movie	Travel
Sentences	2416	2500	1593
Vocabulary	433	583	764
Bigrams	256	368	278
Trigrams	334	499	240

Table 1: Comparison of the three domains: *Carmen*, *Movie*, and *Travel*.

		CARMEN concepts			
TRAVEL concepts	<=>	CITY	GREET	WANT	YES
	CARDNL	5.515	5.457	5.870	5.180
	CITY	**2.719**	**3.146**	3.236	**2.922**
	MONTH	5.595	5.686	6.033	5.514
	WANT	3.335	2.504	**0.914**	2.452
	W.DAY	4.405	4.517	5.080	4.326
	YES	3.233	2.430	3.432	**2.089**

Table 2: Comparison of hand-selected concepts for the *Travel* and *Carmen* tasks.

Three domains were studied: the Carmen-Sandiego computer game, a movie information retrieval service, and a travel reservation system. The corpora were small, on the order of 2500 or fewer sentences. These three domains are compared in Table 1. The set size for each feature is shown; bigrams and trigrams are only included for extant word sequences.

The *Carmen* domain is a corpus collected from a Wizard of Oz study for children playing the Carmen-Sandiego computer game. The vocabulary is limited; sentences are concentrated around a few basic requests and commands. The *Movie* domain is a collection of open-ended questions from adults but of a limited nature, focusing on movie titles, show times, and names of theaters and cities. At an understanding level, the most challenging domain is *Travel*. This corpus is similar to the ATIS corpus, composed of natural speech used for making flight, car and hotel reservations. The vocabulary, sentence structures, and tasks are much more diverse than in the other two domains.

As an initial baseline test of the validity of our proposed metrics, we calculate the KL distances for the *Travel* and *Carmen* domains using hand-selected semantic classes. A concept was used only if there were at least 15 tokens in that class in the domain's corpus. The bigram language model was built using the CMU-Cambridge Statistical Language Modeling Toolkit. Witten Bell discounting was applied and out-of-vocabulary words were mapped to the label UNK. The "backwards LM" probabilities $p^L_a(v|C_{am})$ for the sequences ... vC_{am} ... were calculated by reversing the word order in the training set.

Table 2 shows the symmetric KL distances from the concept-comparison method for a few representative concepts. The minimum distances are in bold for cases where the difference is less than 4 and more than 15% from the next lowest KL distance and multiple entries within 15% are in bold.

Three of the concepts shown here are shared by both domains, <CITY>, <WANT>, and <YES>. The <CITY>, <WANT> and <YES> concepts have the expected KL minima, but <CITY>, <GREET>, and <YES> appear to be confused with each other in the *Carmen* task. This occurs because people frequently used these words by themselves. In addition, children participating in the *Carmen* task frequently prefaced a <WANT> query with the words "hello" or "yes", so that <GREET> and <YES> were used interchangeably. The <CARDINAL> (numbers) and <MONTH> concepts are specific to *Travel* and they have KL distances above 5 for all concepts in the *Carmen* domain. The <W.DAY> category has some similarity to the four *Carmen* classes because people frequently said single-word sentences such as: "hello," "yes", "Monday", or "Boston".

Table 3 shows the KL distances when the concepts in the *Travel* domain are projected into the other two domains, *Car-*

Travel: =>	Carmen	Movie
CARDINAL	4.139	9.931
CITY	**2.718**	**4.174**
DAYPERIOD	4.498	4.542
FIRSTNAME	36.103	40.071
HOTEL	8.790	9.916
MONTH	4.277	19.209
ORDINAL	2.988	4.994
STATE	7.388	9.383
TRAVEL	15.832	6.110
WANT	**1.093**	**1.766**
W.DAY	10.423	9.346
YES	**2.138**	**3.417**

Table 3: Projection of hand-selected concepts from the *Travel* domain to the *Movie* and *Carmen* domains.

men and *Movie*. In this case, each domain's corpus is first parsed only for the words w_{am} that are mapped to the C_{am} concept being projected. Then the right and left bigram LMs for the two domains are calculated. The results show that the ranking is the same for both domains for the three highlighted concepts: <WANT>, <YES>, <CITY>.

Note that for the *Travel* <=> *Carmen* comparisons, the projected distances (Table 3) are almost the same as the compared distances (Table 2) for these three highlighted classes. This suggests these concepts are domain independent and could be used as prior knowledge to bootstrap the automatic generation of semantic classes in new domains [8]. The most common phrases in these three classes are shown for each domain in Table 4 (the hyphens indicate no other phrases commonly occurred). The <WANT> concept is the most domain-independent since people ask for things in a similar way. The <CITY> class is composed of different sets of cities, but they are encountered in similar lexical contexts so the KL distances are small. The sets of phrases in the respective <YES> classes are similar, but they also share a similarity (see Table 2 above) to members of a semantically different class, <GREET>. The small KL distances between these two classes indicates there are some concepts that are semantically quite different, yet tend to be used similarly by people in natural speech. Therefore, the comparison and projection methodologies also identify similarities between groups of phrases based on how they are used by people in natural speech, and not according to their definitions in standard lexicons.

Class	Carmen	Movie	Travel
WANT	I'd like	I would like	I'd like
	I would like	-	I need
	I want	-	I'll need
	-	-	I want
YES	okay	okay	okay
	yeah	yes	yes
	good	fine	that's fine
	-	yeah	yeah
CITY	Alabama	Centerville	Pittsburgh
	Idaho	Warrenville	Boston
	Iowa	Aurora	Cleveland
	New Jersey	CITY theater	Youngstown

Table 4: A comparison of the most common phrases used in the three semantic classes, <WANT>, <YES>, and <CITY>, for all three domains.

Future work will address such issues as evaluating the automatic derivation of semantic classes [9, 10], the soft classification of words and phrases to multiple concepts, and the development of a task hierarchy. Using these techniques, a designer would be able to collect some training sentences, use known statistical techniques to automatically generate semantic classes, and then import additional classes from previously studied tasks that are identified to be similar.

We conclude that both our proposed formalisms for comparing concepts across domains are good measures for ranking concepts according to their degree of domain independence. These metrics could form an extremely powerful tool with which to build understanding modules for new domains.

4. References

[1] S. Nakagawa, "Architecture and Evaluation for Spoken Dialogue Systems," Proc. 1998 Intl. Symp. on Spoken Dialogue, pp. 1-8, Sydney, 1998.

[2] A. Pargellis, H.-K. J. Kuo, C.-H. Lee, "Automatic Dialogue Generator Creates User Defined Applications," Proc. of the Sixth European Conf. on Speech Comm. and Tech., 3:1175-1178, Budapest, 1999.

[3] J. Chu-Carroll, B. Carpenter, "Dialogue management in vector-based call routing," Proc. ACL and COLING, Montreal, pp. 256-262, 1998.

[4] A. N. Pargellis, A. Potamianos, "Cross-Domain Classification using Generalized Domain Acts," Proc. Sixth Intl. Conf. on Spoken Lang. Proc., Beijing, 3:502-505, 2000.

[5] D. Jurafsky et al., "Automatic Detection of Discourse Structure for Speech Recognition and Understanding," Proc. IEEE Workshop on Speech Recog. and Underst., Santa Barbara, 1997.

[6] M. K. McCandless, J. R. Glass, "Empirical acquisition of word and phrase classes in the ATIS domain," Proc. of the Third European Conf. on Speech Comm. and Tech., pp. 981-984, Berlin, 1993.

[7] A. Gorin, G. Riccardi, J. H. Wright, "How May I Help You?," Speech Communications, 23:113-127, 1997.

[8] K. Arai, J. H. Wright, G. Riccardi, A. L. Gorin, "Grammar Fragment Acquisition using Syntactic and Semantic Clustering," Proc. Fifth Intl. Conf. on Spoken Lang. Proc., 5:2051-2054, Sydney, 1998.

[9] K.-C. Siu, H. M. Meng, "Semi-automatic Acquisition of Domain-Specific Semantic Structures," Proc. of the Sixth European Conf. on Speech Comm. and Tech., 5:2039-2042, Budapest, 1999.

[10] E. Fosler-Lussier, H.-K. J. Kuo, "Using Semantic Class Information for Rapid Development of Language Models within ASR Dialogue Systems," Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Salt Lake City, 2001.

[11] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, Prentice Hall, Upper Saddle River, 2000.

[12] P. F. Brown, et al., "Class-based n-gram models of natural language," Computational Linguistics, 18 (4):467-479, 1992.

[13] J. R. Bellegarda, "A latent semantic analysis framework for large-span language modeling," Proc. of the Fifth European Conf. on Speech Comm. and Tech., pp. 1451-1454, Rhodes, 1997.