VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling

Andrea Di Fabio$^{\diamond\heartsuit}$**, Simone Conia**$^{\diamond}$**, Roberto Navigli**$^{\diamond}$

$^{\diamond}$Department of Computer Science
$^{\heartsuit}$Department of Literature and Modern Cultures
Sapienza University of Rome, Italy
`{difabio,conia,navigli}@di.uniroma1.it`

Abstract

We present VerbAtlas, a new, hand-crafted lexical-semantic resource whose goal is to bring together all verbal synsets from Word-Net into semantically-coherent frames. The frames define a common, prototypical argument structure while at the same time providing new concept-specific information. In contrast to PropBank, which defines enumerative semantic roles, VerbAtlas comes with an explicit, cross-frame set of semantic roles linked to selectional preferences expressed in terms of WordNet synsets, and is the first resource enriched with semantic information about implicit, shadow, and default arguments. We demonstrate the effectiveness of VerbAtlas in the task of dependency-based Semantic Role Labeling and show how its integration into a high-performance system leads to improvements on both the in-domain and out-of-domain test sets of CoNLL-2009. VerbAtlas is available at `http://verbatlas.org`.

1 Introduction

During the last two decades, we have witnessed increased attention to Natural Language Understanding, a core goal of Natural Language Processing (NLP). Several challenges, however, are yet to be overcome when it comes to performing sentence-level semantic tasks (Navigli, 2018). In order to understand the meaning of sentences, the semantics of verbs plays a crucial role, since verbs define the argument structure roughly in terms of "who" did "what" to "whom", with the arguments being the constituents that bear a semantic relation (called semantic role) with the verb. In the following example, "Joe" and "lunch" are arguments of "eat", whose argument structure identifies them, respectively, as the Agent and the Patient of the scenario evoked by the verb:

[Joe]$_{\texttt{Agent}}$ is [eating]$_{\texttt{Verb}}$ his [lunch]$_{\texttt{Patient}}$

The automatic identification and labeling of argument structures is a task pioneered by Gildea and Jurafsky (2002) called Semantic Role Labeling (SRL). SRL has become very popular thanks to its integration into other related NLP tasks such as machine translation (Liu and Gildea, 2010), visual semantic role labeling (Silberer and Pinkal, 2018) and information extraction (Bastianelli et al., 2013).

In order to be performed, SRL requires the following core elements: 1) a verb inventory, and 2) a semantic role inventory. However, the current verb inventories used for this task, such as Prop-Bank (Palmer et al., 2005) and FrameNet (Baker et al., 1998), are language-specific and lack high-quality interoperability with existing knowledge bases. Furthermore, such resources provide low to medium coverage of the verbal lexicon (cf. Table 1), with PropBank showing the best figures, but still lower than other lexical inventories like WordNet (Fellbaum et al., 1998). Finally, the informativeness of the semantic roles defined in the various resources ranges from underspecified, as in PropBank's roles, to overspecified, as in FrameNet's frame elements. This poses multiple issues in terms of interpretability or cross-domain applicability.

To overcome the above limitations, in this paper we present VerbAtlas, a manually-crafted inventory of verbs and argument structures which provides several contributions: 1) full coverage of the English verbal lexicon, 2) prototypical argument structures for each cluster of synsets that define a semantically-coherent frame, 3) cross-domain explicit semantic roles, 4) the specification of refined semantic information and selectional preferences for the argument structure of frames, 5) linkage to WordNet and, as a result, to BabelNet (Navigli and Ponzetto, 2010) and Open Multilingual Word-net (Bond and Foster, 2013), which in turn enable scalability across languages. Furthermore, to

	Cluster types	**#**	**Argument roles**	**#**	**Meaning units**	**#**
FrameNet	Frames	1,224	Frame elements	10,542	Lexical units	5,200
VerbNet	Levin's classes	329	Thematic roles	39	Senses	6,791
PropBank	Verbs	5,649	Proto-roles	6	Framesets	10,687
WordNet	–	–	–	–	Synsets	13,767
VerbAtlas	Frames	466	Semantic roles	25	Synsets	13,767

Table 1: Quantitative analysis of popular verbal resources. Since each of these resources is based on different linguistic assumptions, in the top row we use general labels to encompass resource-specific terms that can be viewed as homologous: cluster type, argument roles and meaning units. WordNet does not provide argument structures except for sentence frames which, however, are syntactic and do not specify any roleset.

make VerbAtlas suitable for NLP tasks that rely on PropBank, we also provide a mapping to its framesets. Finally, we prove through an SRL experiment that VerbAtlas is robust and enables state-of-the-art performances on the CoNLL-2009 dataset.

2 Related work

The most popular English verbal resources are FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), and VerbNet (Kipper-Schuler, 2005). Each resource is based on a different linguistic theory, which leads to different information being provided for each verb (cf. Table 1). FrameNet, in particular, was the first resource to be used for SRL (Gildea and Jurafsky, 2002): it is based on frame semantics, theorized by Fillmore (1976), which assumes different roles, i.e., frame elements, for different frames[1]. This led to a proliferation of thousands of roles for only 5200 verbs. Such domain specificity makes it difficult to scale to open-text SRL (Hartmann et al., 2017).

PropBank challenges the issue of FrameNet's roles with a repository of only 6 different core roles plus 19 modifiers for 10,687 framesets[2]. This resource is the most widely adopted for SRL, as also attested by the popularity of datasets such as CoNLL-2005 (Carreras and Màrquez, 2005) and CoNLL-2009 (Hajič et al., 2009). PropBank's methodology was also used for other languages, such as Arabic (Palmer et al., 2008), Chinese (Xue and Palmer, 2003), Spanish and Catalan (Taulé et al., 2008), Hindi-Urdu (Bhatt et al., 2009), Brazilian Portuguese (Duran and Aluísio, 2011), Finnish (Haverinen et al., 2015), Turkish (Şahin and Adalı, 2018), Basque (Aldezabal et al., 2010),

among others. Its application goes well beyond the annotation of corpora: in fact, it was also adopted for the Abstract Meaning Representation (Banarescu et al., 2013), a semantic language that aims at abstracting away from cross-lingual syntactic idiosyncrasies, and NomBank (Meyers et al., 2004), a resource which provides argument structures for nouns. However, PropBank's major drawback is that its roles do not explicitly mark the type of semantic relation with the verb, instead they just enumerate the arguments (i.e., `Arg0`, `Arg1`, etc.). Due to this, role labels do not preserve the same type of semantic relation across verbs, e.g., the first arguments of "eat" and "feel" are both labeled with `Arg0` even if they express different relations (`Agent` and `Experiencer`, respectively).

VerbNet addresses this limit by providing explicit, human-readable roles such as `Agent`, `Patient`, `Experiencer`, etc. Yet, VerbNet suffers from low coverage, in that it includes only 6791 verbs, which makes it a suboptimal resource for wide-coverage SRL. Another drawback of VerbNet is its organization into Levin's classes (Levin, 1993), namely, 329 groups of verbs sharing the same syntactic behavior, independently of their meaning. As a consequence, its classes cannot be used straightforwardly in a semantic task.

On top of their individual limitations, the above resources also have some common drawbacks. One of these is language specificity, which implies a considerable amount of work will be needed for the creation of a corresponding resource for each new language of interest. Another common problem is the lack of coverage, as shown in Table 1. The highest-coverage inventory is PropBank, with its 10,687 framesets and 5,649 verbs. However, PropBank's coverage is still limited when compared to computational lexicons like Word-Net, which contains 13,767 verbal concepts and 11,529 distinct verbs.

---

[1]"By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits." (Fillmore, 1982).

[2]"A frameset corresponds to a coarse-grained sense of the verb which has a specific set of semantic arguments." (Babko-Malaya, 2005).
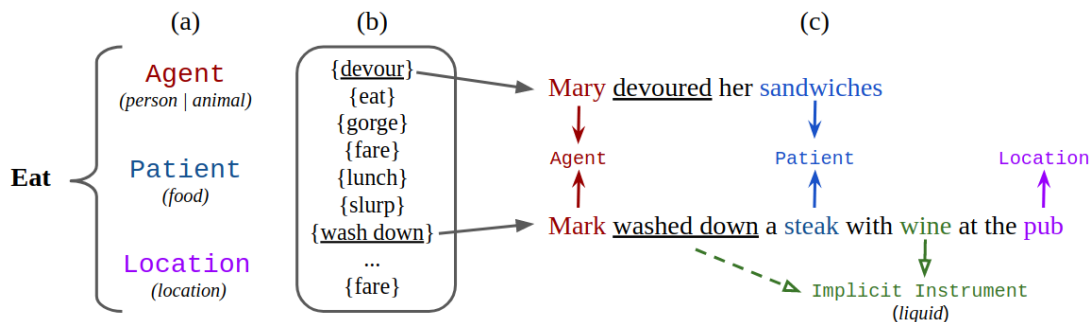
Figure 1: Structure of the EAT frame. The colored roles on the left (a) constitute the Predicate Argument Structure of the frame and their respective selectional preferences within round brackets. Box (b) shows a sample of the frame's synsets (a synonym per synset is shown). In the example sentences (c), we show the head words colored according to their semantic roles. While in the first sentence the synset {devour} does not project the Location, the synset {wash down} projects an implicit Instrument with its synset-level selectional preference *(liquid)*.

To get the best of the three worlds, there have been attempts to map the aforementioned resources to each other. To our knowledge the most popular endeavors are SemLink (Palmer, 2009; Bonial et al., 2013) and Predicate Matrix (De Lacalle et al., 2014), the latter being an extension of the former via automatic methods. However, while the main drawback of SemLink is coverage, the Predicate Matrix suffers from quality issues.

Both the foregoing limitations are addressed in VerbAtlas, the manually-curated resource that we present in this paper. With VerbAtlas we provide the community with a resource that improves the main features of the existing inventories of verbs, while also adding new semantic information. Compared to FrameNet, VerbAtlas has fewer frames and roles while at the same time presenting full coverage in terms of concepts (cf. Table 1). VerbNet's roles, instead, provided inspiration for our role repository, but we reduced the number of roles from 39 to 25 in order to alleviate data sparsity issues. PropBank was mapped to VerbAtlas synsets and roles in order to enable SRL systems to exploit its additional semantic information and improve verb coverage. Moreover, in contrast to the other resources, the use of Word-Net synsets makes VerbAtlas able to scale multilingually through resources such as BabelNet.

In Section 3 we introduce VerbAtlas and its features; in Section 4 we explain how this resource was built and organized. Finally, Section 5 validates VerbAtlas experimentally in the SRL task.

3 VerbAtlas

We now introduce VerbAtlas, a new verbal semantic resource structured into frames which group semantically-coherent synsets from Word-Net (v3.0). A VerbAtlas frame is a cluster of verb meanings expressing similar semantics, which expands upon the frame notion of FrameNet. Each frame is provided with an argument structure that generalizes over all the synsets in the frame plus preferential selections for each semantic role. Furthermore, synsets are enriched with novel semantic information. In Figure 1 we show the structure of the EAT frame in VerbAtlas, which we use as a running example to illustrate the new features in our work and compare them with current verbal resources.

3.1 Frame organization

We define a frame in VerbAtlas as a cluster of WordNet synsets that, with different shades of meaning, express a certain scenario. For instance, the EAT frame, an excerpt of which is shown in Figure 1(b), comprises all the synsets specializing the general scenario of "eating", including synsets such as {eat}, {devour, guttle, raven, pig}, etc.

This organization of frames is intended to overcome the limitations affecting VerbNet and FrameNet. In fact, while the former organizes verbs by syntactic rather than semantic behavior, the latter is affected by the sparsity of 5,200 verbal senses (i.e., lexical units) distributed across frames. Instead, PropBank's framesets correspond to a coarse-grained sense of a verb, but each frameset expresses its own separate argument structure.

Differently from the other resources, VerbAtlas frames are organized into 466 wide-coverage and semantically-coherent clusters (cf. Table 1) which provide cross-frame argument structures.

3.2 Semantic roles

A limit in PropBank is its inventory of so-called proto-roles (`Arg0`, `Arg1`, etc., with the first two roughly corresponding, respectively, to a proto-agent and a proto-patient (Dowty, 1991)), which does not provide human-readable labels and is predicate-specific (i.e., the same label does not necessarily depict the same semantic relation across predicates).

While PropBank roles do not provide clear explicit semantics, FrameNet roles are explicit but frame-specific (e.g., `Ingestor` and `Ingestibles` for the "Ingestion" frame, or `Impactor` and `Impactee` for the "Impact" frame). This produces a fine-grained role inventory that makes it difficult for SRL systems to generalize across frames. In VerbAtlas we follow VerbNet and take a middle-ground approach: our inventory of 25 semantic roles is inspired by VerbNet, whose 39 labels (like `Agent`, `Patient`, `Time`, etc.) are explicit, cross-frame and domain-general. The rationale is that these features enable neural networks employed in the SRL task to generalize across frames in a consistent way, as we show in Section 5.2. For example, we can tag the arguments of different VerbAtlas frames – such as EAT, HIT and CONQUER – with just two roles, namely: `Agent` and `Patient`.

3.3 Prototypical Argument Structure

Each VerbAtlas frame expresses a Prototypical Argument Structure (PAS) that generalizes over all the synsets in a particular frame. The PAS specifies a roleset that defines the frame's overall meaning (e.g., in Figure 1(a) `Agent`, `Patient` and `Location`). In Figure 1(c) we show two sentences annotated with argument roles.

Since we do not distinguish between core roles and adjuncts, we decided to also include in the various PAS roles which might be projected optionally by argument structures that are nonetheless present in the scenario evoked by the frame. For instance, the inclusion of the `Location` role in the PAS of the EAT example ensures a robust descriptiveness of the PAS across synsets in the same frame.

3.4 Selectional preferences

To narrow down the number of candidates for a particular argument slot and provide further semantic structure, the semantic role of the PAS has been manually labeled with selectional preferences from a set of 116 macro-concepts. Our selectional preferences are defined by WordNet synsets whose hyponyms are expected to be likely candidates to the corresponding argument slot, a strategy similar to that of Agirre and Martinez (2002) which, instead, was algorithm-based.

Consider again the EAT frame and its PAS: `Agent`, `Patient`, `Location` (Figure 1 (a)). In this frame, most of the example sentences from the WordNet synsets express a `Patient` like "cake", "meat" or "banana", thus, since their common hypernym is "food", we provided the PAS with the information that the `Patient` prototypically expects hyponyms of the {food, solid food} synset. The `Location` role is labeled with its homonymous {location} synset given its generality.

3.5 Synset-level semantic information

To enrich the semantic representation of synsets, VerbAtlas provides semantic and pragmatic (English-specific) information regarding implicit (1,028 labels), shadow and default arguments (2,979 labels) inspired by Pustejovsky (1995) and inferred from synset's glosses and examples. To our knowledge, the following information is new in a large-scale verbal resource such as ours:

- **implicit arguments**, i.e., implicit in the argument structure of the verb but not always syntactically expressed. Consider the synset {overleap, vault} in our JUMP frame: since its gloss is "Jump across or leap over (an obstacle)", we know that the `Patient` of this verb can be a hyponym of {obstacle}, therefore implying a selectional preference on the role with the {obstacle} synset.

- **shadow and default arguments**: the former is incorporated in the meaning of the verb but not syntactically expressed. An example from the EAT frame is {eat in, dine in} ("Eat at home"). This synset is tagged with the shadow argument `Location` = {home}, since the latter is not expressed syntactically. On the other hand, default arguments are logically implied but not syntactically expressed. These are also tagged as shadow arguments. For instance, the synset {deliver} (as in "Our local supermarket delivers") has the label `Patient` = {grocery} to provide the commonsense information that what a supermarket usually delivers is groceries.

We are aware of the problems raised about Pustejovsky's framework of analysis (Fodor and Lepore, 1998), but we believe that this new information, if properly exploited, would be fruitful for better meaning representations.

3.6 Linkage to PropBank and multilingual resources

To allow a straightforward applicability to and evaluation of VerbAtlas on semantic tasks based on PropBank, such as SRL, each frameset and roleset in the CoNLL-2009 dataset was mapped to the corresponding VerbAtlas frame and roleset. This mapping was done starting from the Predicate Matrix (De Lacalle et al., 2014) and then manually corrected and augmented. Finally, we aligned PropBank's roles with the corresponding roles in VerbAtlas PAS (e.g., `Arg0` and `Arg1` of PropBank's eat.01 frameset were mapped, respectively, to the VerbAtlas PAS roles `Agent` and `Patient` of the frame EAT).

Moreover, thanks to the use of WordNet and its semantic nature, VerbAtlas can easily scale to arbitrary languages. This can be achieved by leveraging BabelNet (Navigli and Ponzetto, 2010), a lexical-semantic resource that provides multilingual synsets in 284 different languages linked to WordNet itself. As a result, VerbAtlas can be used in virtually any language, in contrast to PropBank and other resources which are inherently language-specific and require considerable human intervention in each new language.

Finally, the above implies that if there is a verbal repository for one of the languages linked to the aforementioned resources, its argument structures can be seamlessly aligned with the PAS of VerbAtlas, as well as with VerbAtlas frames.

4 Methodology

4.1 Bottom-up approach

The manual construction of VerbAtlas was performed in a bottom-up fashion. Rather than forcing synsets into a predetermined set of frames, we started the clustering process from the full inventory of 13,767 WordNet synsets, which represent the building blocks of VerbAtlas frames. This allowed us to induce the semantics of 466 frames (with an average of 29.5 synsets per frame) and make VerbAtlas consistent both in terms of lexical semantics and the frame's argument structures.

EAT	CLASSIFY	SWITCH
KILL	CLEAN	MOVE SOMETHING
DRINK	AMELIORATE	JUDGE
POUR	TRANSPORT	PERFORM
OBTAIN	BEHAVE	STOP
GIVE	HIT	COMBINE

Table 2: Frames with the highest number of synsets.

4.2 Creation of frames

VerbAtlas frames were induced via semantic similarity between synsets. During multiple iterations over the verb inventory, if two or more synsets were perceived as similar, namely, if they shared features like the purpose of the action and the participants in the action (Hill et al., 2015) (e.g., "dine" and "lunch"), they were clustered together to form a new semantically-coherent frame. A one-synset-one-frame strategy was used to avoid any future mapping problem with other resources. The process is based on synset-by-synset human inspection. Two synsets are clustered together in the same frame if they express semantically-similar scenarios. For example, the verbs "kill" and "slaughter" share similar participants in the action (an `Agent` who kills/slaughters; a `Patient` who is killed/slaughtered) and purpose (`Agent` makes `Patient` die), so they are clustered into the KILL frame.

At the end of the first iteration, we checked the resulting frames and named them according to the common action implied by the synsets contained therein. For example, the EAT frame (Figure 1(b)) is composed of synsets that depict different kinds of action implying eating, like {devour, ..., pig} and {gorge, ..., glut}. In Table 2 we report the frames with the highest number of verb synsets.

To validate the resulting frames we adopted a strategy similar to Hovy et al. (2006): we provided 3 linguists not involved in the clustering with a random sample of 1000 frame-synset pairs and asked them if the action expressed by the synset was implied by that expressed by the frame. We iterated over the inventory various times and moved the synsets from one frame to another until the Cohen's Kappa coefficient (Eugenio and Glass, 2004) of their (yes or no) agreement was $\kappa \geq 0.80$. Once this value was attained, we finalized the overall clustering and the resulting frames.

4.3 Establishing explicit semantic roles

Inspired by existing research on semantic roles (Bonial et al., 2011; Allen and Teng, 2018),

Agent	Material
Attribute	Patient
Beneficiary	Product
Cause	Purpose
Co-Agent	Recipient
Co-Patient	Result
Co-Theme	Source
Destination	Stimulus
Experiencer	Theme
Extent	Time
Goal	Topic
Instrument	Value
Location	

Table 3: List of VerbNet roles which make up the VerbAtlas role inventory.

Affector	Manner
Axis	Path
Context	Pivot
Duration	Precondition
Final_Time	Predicate
Initial_Location	Reflexive
Initial_State	Trajectory

Table 4: List of VerbNet roles unused in VerbAtlas.

VerbAtlas implements VerbNet's human-readable role labels across its frames and merges together some VerbNet roles which can be seen as complementary. For example, `Initial_Location` and `Initial_State` are subsumed by `Source`. The hunch is that we can consider them as playing the same role but in different scenarios: the former is the `Source` for verbs of change of location, while the latter is the `Source` for verbs of change of state. Furthermore, in VerbAtlas we did not want to use poorly instantiated roles (e.g., `Affector`), therefore 14 of the roles (Table 4) were subsumed by coarser and more common roles (e.g., `Agent` in place of `Affector`, see Table 3).

4.4 Rationale of a Prototypical Argument Structure

We defined as "prototypical" an argument structure capable of being applied to all the synsets in a frame. To achieve this, each PAS was defined at the end of the first iteration over the inventory of synsets and constantly adjusted until the last iteration. The initial PAS was inspired by the argument structure of the common verbal concept implied by the synsets constituting the frame. For instance, in the case of BETRAY, `Agent` and `Patient`. The PAS was later expanded due to synsets inside the frame that projected additional arguments (e.g., the `Goal` role of {defect, desert}

as in the sentence "The reporter defected to another *network*").

4.5 Rationale of selectional preferences

Each semantic role of a PAS was provided with selectional preferences from a set of 116 synsets. The set was created by generalizing across the arguments in the example sentences of each synset in a given frame. The result for each PAS argument was one or more hypernyms in the WordNet taxonomy which were common across the frame synsets.

In contrast to VerbNet selectional restrictions, the selectional preferences in VerbAtlas do not restrict the occurrence of words with a particular feature (e.g., solid, liquid, etc.), rather, they suggest the most prototypical hypernym(s) for a given argument. We opted for preferences instead of restrictions to not exclude the potential metaphorical use of a verb.

5 Experiments

In previous sections we discussed the qualitative and structural advantages of VerbAtlas compared to existing resources. Because, in principle, a new linguistic resource might seem to provide an arbitrary contribution, we also provide here an experimental validation of its usefulness in an SRL task. First, we present our experimental setup (Section 5.1) and then discuss our results (Section 5.2).

5.1 Experimental setup

Goal We argue that the quality and wealth of information in VerbAtlas can effectively improve the performance of an existing, high-performance SRL system like Cai et al. (2018).

Datasets We performed our experiments on the English part of CoNLL-2009 in-domain and out-of-domain datasets (Hajič et al., 2009) designed for the dependency-based SRL task and whose annotations concern predicate argument structures from PropBank and nominal argument structures from NomBank (Meyers et al., 2004). The dataset consists of 39,279 sentences and 958,167 tokens (18.7% of which are argument bearers).

Through this experiment we aim to show that the additional information provided by VerbAtlas frames and semantic roles improves the performance of a neural network on the in-domain test set. We also aim to demonstrate a better ability to generalize on the out-of-domain dataset.
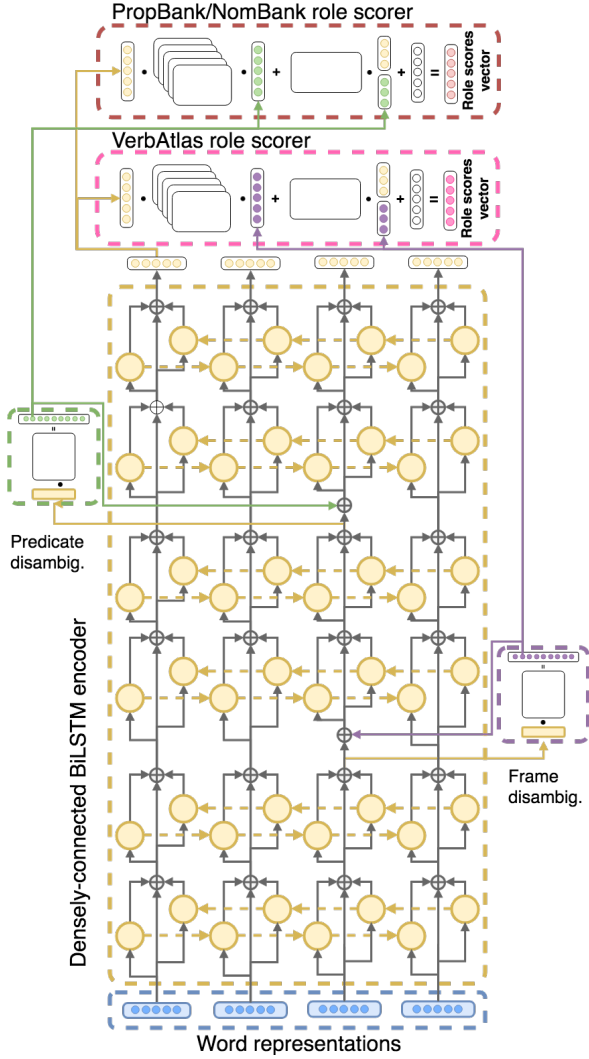
Figure 2: Overview of the model architecture. From bottom to top, a sequence of word embeddings is fed to a densely-connected BiLSTM encoder where the output of each encoding layer is concatenated with its input. The predicate hidden representation from the second BiLSTM layer is used for VerbAtlas frame disambiguation (right), whereas the predicate hidden representation from the fourth BiLSTM layer is used for PropBank predicate disambiguation (left). The topmost output of the BiLSTM encoder is used together with the output of the PropBank predicate disambiguation layer to obtain a PropBank-style SRL output, and together with the output of the VerbAtlas frame disambiguation layer to obtain a VerbAtlas-style SRL output.

Baseline model Our baseline model in Figure 2 is built on top of the syntax-agnostic model proposed by Cai et al. (2018) in that it is mainly composed of a word representation layer, a sequence encoder and a biaffine attentional role scorer. A key difference is that our model features a multi-output layer that returns PropBank (PB) and Nom-Bank (NB) labels (i.e., framesets and their roles)

and, if the predicate is a verb, also VerbAtlas labels (i.e., frames and their roles). With this design choice, the output of our model can be directly compared to the output of previous SRL models. At the same time, our model achieves a deeper and more general understanding of the relations between a PB or NB predicate and its arguments by learning from the corresponding VerbAtlas frame and its semantically-coherent roles.

Formally, our model is built on top of the following components:

- A **word representation layer** that, given a sentence $\mathbf{s} = \langle w_1, w_2, \ldots, w_n \rangle$, builds a sequence of word representations $\mathbf{x} = \langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \rangle$ where $\mathbf{x}_i$ is an embedding representing w_i. Each $\mathbf{x}_i$ is the result of the concatenation of a pre-trained word embedding $\mathbf{e}^{pt}$, and the following trainable vectors: a word embedding $\mathbf{e}^w$, a lemma embedding $\mathbf{e}^l$, a POS embedding $\mathbf{e}^{pos}$, and a predicate lemma embedding $\mathbf{e}^{pred}$ (active only if w_i is a predicate). Formally: $\mathbf{x}_i = \mathbf{e}^{pt} \oplus \mathbf{e}^w \oplus \mathbf{e}^l \oplus \mathbf{e}^{pos} \oplus \mathbf{e}^{pred}$. We use GloVe embeddings (Pennington et al., 2014) as our underlying pre-trained word embeddings.

- A **densely-connected BiLSTM encoder** that, given a sequence $\mathbf{x}$ of word representations, returns a sequence of encodings $\mathbf{y} = \langle \mathbf{y}_i = \text{BiLSTM}(\mathbf{x}_i; \mathbf{x}) : \forall i \in \{1, \ldots, n\} \rangle$, where $\mathbf{y}_i$ is a dynamic representation of $\mathbf{x}_i$ with respect to the context defined by the whole sequence $\mathbf{x}$. In a densely-connected BiLSTM encoder, the output of each layer is concatenated with the input of the same layer to mitigate the vanishing gradient problem. If $\mathbf{h}_i^k$ is the encoding of the k-th layer for $\mathbf{x}_i$, then $\mathbf{h}_i^{k+1} = \mathbf{h}_i^k \oplus \text{LSTM}^f(\mathbf{h}_i^k; \mathbf{h}_{1:i-1}^k) \oplus \text{LSTM}^b(\mathbf{h}_i^k; \mathbf{h}_{i+1:n}^k)$, and $\mathbf{y}_i = \mathbf{h}_i^m$ where $m = 6$ is the final BiLSTM layer, while LSTM^f and LSTM^b are the forward and backward LSTM transformations.

- A **frame disambiguation layer** that, given the BiLSTM encoding $\mathbf{y}_{pred}^i$ of a predicate w_{pred} at the i-th encoder layer (with $i = 2$), disambiguates w_{pred} with a VerbAtlas frame f, returning a trainable frame embedding $\mathbf{e}^f$:

$$\mathbf{h}_{pred}^i = \text{ReLU}(\mathbf{W}' \cdot \mathbf{y}_{pred}^i + \mathbf{b}')$$

$$f = \text{argmax}(\mathbf{W}^f \cdot \mathbf{h}_{pred}^i + \mathbf{b}^f)$$

- A **predicate disambiguation layer** that, given the BiLSTM encoding $\mathbf{y}_{pred}^j$ of a predicate w_{pred} at the j-th encoder layer (with $j = 4$) and the frame embedding $\mathbf{e}^f$, disambiguates w_{pred} with a PB or NB frameset p, returning a trainable predicate embedding $\mathbf{e}^p$:

$$\mathbf{h}_{pred}^j = \mathrm{ReLU}(\mathbf{W}'' \cdot (\mathbf{y}_{pred}^j \oplus \mathbf{e}^f) + \mathbf{b}'')$$

$$p = \mathrm{argmax}(\mathbf{W}^p \cdot \mathbf{h}_{pred}^j + \mathbf{b}^p)$$

- A **biaffine attentional PB and NB role scorer** that, given a BiLSTM encoding $\mathbf{y}_i$ for a word w_i and a predicate embedding $\mathbf{e}^p$, returns a vector $\mathbf{s}_i^p$ of PB/NB role scores for w_i with respect to p in a similar fashion to Cai et al. (2018). Formally:

$$\mathbf{s}_i^p = \mathbf{y}_i^\top \cdot \mathbf{W}_{\mathrm{PB}}^{role} \cdot \mathbf{e}^p + \mathbf{U}_{\mathrm{PB}}^{role}(\mathbf{y}_i \oplus \mathbf{e}^p) + \mathbf{b}_{\mathrm{PB}}^{role}$$

- A **biaffine attentional VerbAtlas role scorer** that, given a BiLSTM encoding $\mathbf{y}_i$ for a word w_i and a frame embedding $\mathbf{e}^f$, returns a vector $\mathbf{s}_i^f$ of VerbAtlas role scores for w_i with respect to f. Formally:

$$\mathbf{s}_i^f = \mathbf{y}_i^\top \cdot \mathbf{W}_{\mathrm{VA}}^{role} \cdot \mathbf{e}^f + \mathbf{U}_{\mathrm{VA}}^{role}(\mathbf{y}_i \oplus \mathbf{e}^f) + \mathbf{b}_{\mathrm{VA}}^{role}$$

We found it beneficial to interleave the frame disambiguation layer and the predicate disambiguation layer within the BiLSTM encoder layers, enhancing the input of the upper encoder layers with the corresponding frame and predicate embeddings. Finally, our model loss is defined as:

$$l_{\mathrm{total}} = l_{\text{ PropBank/NomBank roles}} + l_{\text{ VerbAtlas roles}} +$$
$$l_{\text{ predicate disambiguation}} + l_{\text{ frame disambiguation}}$$

Since the choice of hyperparameters for an LSTM-based model can significantly affect results (Reimers and Gurevych, 2017), we trained our baseline using the same values as in Cai et al. (2018) unless otherwise stated.

5.2 Results

In-domain SRL Table 5 reports the results of our syntax-agnostic baseline model on the CoNLL-2009 in-domain test set. Not only does our model outperform the syntax-agnostic model of Cai et al. (2018) by a significant margin (according to a χ^2 test, p < 0.05, on precision and recall) with a 0.4% F_1 improvement, but it also slightly outperforms the syntax-aware model of Li

Syntax-aware system	P	R	F_1
Roth and Lapata (2016)	88.1	85.3	86.7
Marcheggiani and Titov (2017)	89.1	86.8	88.0
He et al. (2018)	89.7	**89.3**	89.5
Li et al. (2018)	**90.3**	**89.3**	**89.8**
Syntax-agnostic system	P	R	F_1
Marcheggiani et al. (2017)	88.7	86.8	87.7
He et al. (2018)	89.5	87.9	88.7
Cai et al. (2018)	89.9	89.2	89.6
This work	**90.5**	**89.5**	**90.0**

Table 5: Results on the English in-domain test.

Syntax-agnostic system	P
He et al. (2018)	95.5
Cai et al. (2018)	95.0
This work	**96.0**

Table 6: Predicate disambiguation results on the English in-domain test. In CoNLL-2009 predicates are already identified, hence we only report predicate disambiguation precision.

Syntax-agnostic system	P	R	F_1
Marcheggiani et al. (2017)	79.4	76.2	77.7
He et al. (2018)	**81.7**	76.1	78.8
Cai et al. (2018)	79.8	78.3	79.0
This work	81.1	**78.4**	**79.7**

Table 7: Results on the English out-of-domain test.

et al. (2018) with a 0.2% F_1 improvement. We recall that the main difference between our baseline model and that of Cai et al. (2018) lies in the additional VerbAtlas frame disambiguation layer and role scorer. This demonstrates the boost provided by VerbAtlas in PropBank/NomBank-based SRL.

In-domain predicate disambiguation Remarkably, as shown in Table 6, our model outperforms the previously reported best scores in the predicate disambiguation subtask, scoring 96.0% in precision for the in-domain test set and reporting a 1.0% improvement over Cai et al. (2018) and a 0.5% improvement over He et al. (2018).

Out-of-domain SRL The results of our baseline model on the out-of-domain SRL CoNLL-2009 dataset are shown in Table 7. Our system improves the results reported by Cai et al. (2018) by 0.7% F_1, proving that using VerbAtlas improves the model's ability to generalize across domains.

F_1 evaluated on	verbs + nouns	only verbs
This work	90.0	91.1
without VerbAtlas	89.5	90.5
without PB *nor* NB	–	90.9

Table 8: The model shows a significant F_1 decrease when it is trained without exploiting VerbAtlas frames and roles (second row) falling in line with Cai et al. (2018) (89.5% vs 89.6%, respectively). On the contrary, removing the PropBank role scorer leads to a negligible performance drop (last row).

5.3 Analysis

Our main experiment proved that a semantics-focused resource for verbal predicates such as VerbAtlas can be successfully employed for CoNLL-like SRL datasets that include a mix of nominal and verbal predicates. But what is the contribution of VerbAtlas (i.e., of its semantically-clustered frames and semantically-coherent roles) to the overall performance? How does a VerbAtlas-only model fare when evaluated solely on verbal predicates against a PropBank-only model? To answer these questions, we evaluated the model in two further settings.

A first study aimed at quantifying the boost in performance the model gets from the use of VerbAtlas. Removing the VerbAtlas frame disambiguation layer and role scorer from our model significantly decreases (according to a χ^2 test, p < 0.05, on precision and recall) the overall performance in F_1 score by 0.5%, as reported in Table 8 (left column), with results that are comparable to those of Cai et al. (2018) (89.5% vs 89.6% F_1, Tables 8 and 5, respectively). This proves that the model with VerbAtlas achieves a better understanding of semantic roles, thanks to using the VerbAtlas frame disambiguation layer and role scorer.

A second study aimed at comparing PropBank and VerbAtlas on their common ground, i.e., their coverage and organization of verb meanings into framesets and frames, respectively. Table 8 (right column) shows that removing the PropBank biaffine role scorer leads to a negligible performance drop (0.2% F_1) in how it performs on argument labeling of verb predicates using only the semantic roles of VerbAtlas mapped at the output layer to PropBank roles (see Section 3.6); we note that the higher number of roles in VerbAtlas makes the argument labeling problem potentially harder. In contrast, when VerbAtlas is removed, the drop in performance is noteworthy, with a 0.6% decrease in F_1 on SRL of verbal predicates.

6 Conclusions and Future Work

In this paper we presented VerbAtlas, a new large-scale verbal semantic resource which provides generalizing argument structures with cross-frame semantic roles. The resource is available at `http://verbatlas.org`.

In contrast to other verb repositories, VerbAtlas offers full coverage of the English verbs and addresses the issues of current predicate resources, while at the same time providing linkage to Word-Net and PropBank. This makes the resource fully compatible with previous datasets and scalable to arbitrary languages thanks to BabelNet.

While the frame creation process resulted in a strong agreement between annotators, we further validated the quality of VerbAtlas experimentally by showing that the integration of its frame information together with its explicit semantic roles enables a neural architecture to improve its performance on the Semantic Role Labeling task. This improvement translates across domains, demonstrating the robustness and variety of the knowledge provided in our resource.

As future work, we plan to take full advantage of the novel semantic features available in VerbAtlas, such as wide-coverage selectional preferences and synset-level information, by exploiting them in multilingual SRL and Word Sense Disambiguation tasks. Our plans include integrating the selectional preferences from SyntagNet (Maru et al., 2019), a new, large-scale lexical-semantic combination resource. We also plan to extend our methodology to nouns and adjectives, in a similar fashion to (O'Gorman et al., 2018) and connect the resulting frames to those in VerbAtlas.

References

Eneko Agirre and David Martinez. 2002. Integrating selectional preferences in WordNet. In *Proceedings of First International WordNet Conference. Mysore (India)*, pages 1–9.

Izaskun Aldezabal, María Jesús Aranzabe, Arantza Díaz de Ilarraza Sánchez, and Ainara Estarrona. 2010. Building the Basque PropBank. In *Proceedings of LREC*.

James Allen and Choh Man Teng. 2018. Putting semantics into semantic roles. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 235–244.

Olga Babko-Malaya. 2005. PropBank annotation guidelines. *URL: http://verbs.colorado.edu*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 86–90. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2013. Textual inference and meaning representation in human robot interaction. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 65–69.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for Hindi-Urdu. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 186–189.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1352–1362.

Claire Bonial, William Corvey, Martha Palmer, Volha V Petukhova, and Harry Bunt. 2011. A hierarchical unification of LIRICS and verbnet semantic roles. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 483–489. IEEE.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9–17.

Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2753–2765.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 152–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maddalen Lopez De Lacalle, Egoitz Laparra, and German Rigau. 2014. Predicate Matrix: extending SemLink through WordNet mappings. In *Proceedings of LREC*, pages 903–909.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Magali Sanches Duran and Sandra Maria Aluísio. 2011. PropBank-Br: a Brazilian Portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

Christiane Fellbaum et al. 1998. *WordNet: An electronic database*. MIT Press, Cambridge, MA.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Charles J. Fillmore. 1982. Frame semantics. *Linguistics in the Morning Calm*, pages 111–137.

Jerry A. Fodor and Ernie Lepore. 1998. The emptiness of the lexicon: reflections on James Pustejovsky's The Generative Lexicon. *Linguistic Inquiry*, 29(2):269–288.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482.

Katri Haverinen, Jenna Kanerva, Samuel Kohonen, Anna Missilä, Stina Ojala, Timo Viljanen, Veronika Laippala, and Filip Ginter. 2015. The Finnish proposition bank. *Language Resources and Evaluation*, 49(4):907–926.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2061–2071.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pensylvania.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. 2018. A unified syntax-aware framework for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2411.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 716–724.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In *Proceedings of EMNLP-IJCNLP 2019*.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*.

Roberto Navigli. 2018. Natural Language Understanding: Instructions for (present and future) use. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pages 5697–5702, Stockholm, Sweden.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. Association for Computational Linguistics.

Tim O'Gorman, Sameer Pradhan, Martha Palmer, Julia Bonn, Kathryn Conger, and James Gung. 2018. The new Propbank: Aligning Propbank with AMR through POS unification. In *Proceedings of LREC*.

Martha Palmer. 2009. Semlink: Linking Propbank, VerbNet and FrameNet. In *Proceedings of the generative lexicon conference*, pages 9–15. GenLex-09, Pisa, Italy.

Martha Palmer, Olga Babko-Malaya, Ann Bies, Mona T. Diab, Mohamed Maamouri, Aous Mansouri, and Wajdi Zaghouani. 2008. A pilot arabic Propbank. In *LREC*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543. ACL.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge MA.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 338–348.

Gözde Gül Şahin and Eşref Adalı. 2018. Annotation of semantic roles for the Turkish proposition bank. *Language Resources and Evaluation*, 52(3):673–706.

Carina Silberer and Manfred Pinkal. 2018. Grounding semantic roles in images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2616–2626.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of LREC*.

Nianwen Xue and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese Treebank. In *Proceedings of the Second Workshop on Chinese Language Processing, SIGHAN 2003, Sapporo, Japan, July 11-12, 2003*.