Corpus-driven Thematic Hierarchy Induction

Ilia Kuznetsov and **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab (UKP-TUDA) and Research Training Group AIPHES
Department of Computer Science
Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de/

Abstract

Thematic role hierarchy is a linguistic tool used to describe interactions between semantic roles and their syntactic realizations. Despite decades of dedicated research and numerous thematic hierarchy suggestions in the literature, this concept has not been used in NLP so far due to incompatibility and limited scope of existing hierarchies. We introduce an empirical framework for thematic hierarchy induction and evaluate several role ranking strategies on English and German corpus data. We hypothesize that inducing a thematic hierarchy is feasible, that a hierarchy can be induced from small amounts of data and that resulting hierarchies apply cross-lingually. We evaluate these assumptions empirically.

1 Introduction

Semantic roles are one of the core concepts in NLP, and automatic semantic role labeling (SRL) is a major task with applications in question answering (Shen and Lapata, 2007), machine translation (Liu and Gildea, 2010) and information extraction (Christensen et al., 2010). The goal of SRL is to label the semantic arguments of a predicate (e.g. a verb) with roles from a pre-defined role inventory. Conceptually, role assignment in SRL can be split in two steps: local labeling estimates the likelihood of a certain semantic argument bearing a certain role; global optimization takes context-dependent *role interactions* into account and enforces certain theoretically motivated constraints (e.g. "each role must appear only once per predication").

State of the art in SRL is held by the systems based on deep neural networks (Marcheggiani and Titov, 2017; He et al., 2017). While achieving remarkable quality on benchmark datasets, modern systems show a considerable ≈10-point performance drop when applied out-of-domain. This issue is aggravated by the fact that deep neural networks require significant amounts of training data, and SRL annotations are expensive to produce. While local role assignment can be augmented using unlabeled data (e.g. via pre-trained word and character embeddings), context-dependent role interaction is an SRL-specific phenomenon and can only be learned from annotated SRL corpora.

Aiming to reduce the training data requirements for SRL, we revisit the notion of **thematic hierarchy** (TH), a compact *delexicalized* way to model context-dependent role interactions. Thematic hierarchies assume that given a syntactic hierarchy (e.g. *subject* $\prec$[1] *object* $\prec$ *oblique*) semantic roles can be ranked in a way that higher ranked roles take higher ranked syntactic positions. One example of phenomena captured by THs is the choice of subject: given a thematic hierarchy `Agent` $\prec$ ... $\prec$ `Instrument`, an `Instrument` can only become subject if the `Agent` is not present, e.g. *"[John]$_{Ag}$ broke the window with a [hammer]$_{In}$"* $\rightarrow$ *"A [hammer]$_{In}$ broke the window"*.

THs have received considerable attention in linguistic literature, but were so far impractical for use in NLP and SRL due to incompatibility and limited scope of the existing hierarchies. As a first step towards including THs into the NLP tool inventory we suggest an empirical framework for inducing THs from role-annotated corpora. Since VerbNet (Schuler, 2006) is the only SRL framework that operates with thematic roles, we choose it as our basis and perform experiments on the PropBank corpus (Palmer et al., 2005) enriched with VerbNet role labels via SemLink (Bonial et al., 2013).

The contributions of this paper are as follows:

- We suggest a method for global thematic hierarchy induction from corpus data;

[1] We use $\prec$ for rank precedence, and / for ties

Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), pages 54–64
Brussels, Belgium, October 31 - November 1, 2018.

- We propose several thematic and syntactic ranking models and evaluate them on English and German data;
- We show that thematic hierarchies can be induced and applied cross-lingually while leaving room for improvement; we further show that thematic hierarchy induction is data-efficient and can produce a high-quality hierarchy using just a fraction of training data.

2 Related work

2.1 Semantic roles and the Lexicon

Semantic roles in the modern sense have been introduced in 1960s as a way to account for variation in syntactic behavior of verbs which can not be explained by purely syntactic means (Gruber, 1965; Fillmore, 1968). A commonly used motivational example contrasts the use of verbs *hit* and *break*: while both are regular transitive verbs, *hit* does not allow construction (4); and construction (5) is ungrammatical in both cases.

(1) $[\text{John}]_X$ broke/hit the $[\text{window}]_Y$ with a $[\text{stone}]_Z$.
(2) $[\text{John}]_X$ broke/hit the $[\text{window}]_Y$.
(3) A $[\text{stone}]_Z$ broke/hit the $[\text{window}]_Y$.
(4) The $[\text{window}]_Y$ broke/*hit.
(5) The $[\text{window}]_Y$ *broke/*hit with a $[\text{stone}]_Z$.

There exist several principled ways to describe the syntactic behavior of arguments in the lexicon. Available constructions can be defined individually on **verb sense** basis. This strategy is precise but highly redundant, since verbs show substantial similarities in syntactic behavior; besides, it does not generalize to the out-of-vocabulary (OOV) predicates.

A step towards a more general representation is **verb class** grouping (Levin, 1993): verbs senses can be grouped into verb classes with syntactic behavior shared among the members of the class. For example, syntactically *break* behaves like *crash*, *shred* and *split*, while *hit* behaves like *bash* and *whack* in the corresponding verb senses. This significantly reduces the lexicon redundancy and allows treatment of the OOV verbs if the verb class can be determined. A similar level of granularity is used by the major SRL frameworks: FrameNet SRL (Das et al., 2010) and, to some extent, PropBank SRL (Roth and Woodsend, 2014).

Semantic arguments share similarities across verb classes, giving rise to the notion of general semantic roles. While there exists no consensus on the inventory of semantic roles, a subset shared by the most theoretical approaches includes roles such as `Agent` (the active sentient initiator of the event), `Theme` (the most affected participant), `Result` (the outcome of the event), `Instrument` (the instrument used) etc. Semantic roles show similar behavior across languages and can be thought of as grammatically relevant universal categories humans use to conceptualize real-world events. Following common terminology, we further refer to *general*, *predicate-independent* semantic roles as **thematic roles**. This level of granularity is, for example, used by VerbNet (Schuler, 2006).

Thematic roles' syntactic behavior depends on the presence of other thematic roles in the sentence: as our example above demonstrates, an `Instrument` can only take the subject position if the `Agent` is not present (3); and `Theme` can only become subject if both `Agent` and `Instrument` are not expressed (4-5). A widely used modeling tool to account for context dependency is the **thematic hierarchy** (TH): given a syntactic prominence scale (e.g. *subject* $\prec$ *oblique...* $\prec$ *object*), one can assume that there exists a universal ranking of thematic roles, which is homomorphic to the syntactic ranking (e.g. `Agent` $\prec$ `Instrument` $\prec$ `Theme`). The top-ranking semantic argument gets assigned to the highest available syntactic position, the second-ranking gets the second-highest position, etc.

THs are a compact delexicalized way to describe semantic roles' syntactic behavior at the grammar level, which could reduce data requirements and improve generalization capability of SRL systems. However, THs from the literature come from varying theoretical backgrounds, are based on different syntactic formalisms and operate with different role inventories. Most of these THs are justified via basic (often synthetic) language examples, aiming to verify a certain theory cross-lingually rather than to describe the language use in a compact way.

2.2 Major SRL Frameworks

The choice of linguistic theory in SRL is mostly dictated by the availability of training data. PropBank SRL is based on the PropBank corpus (Palmer et al., 2005) which utilizes a set of predicate-specific core roles (`A0-5`) and a set

of general, predicate-independent adjunct roles (AM-TMP, AM-LOC etc.). Core roles are defined on verb sense level. An effort is made to ensure consistency in assigning A0 (Agent-like) and A1 (Patient-like). The rest of the core arguments (A2-5) are verb sense-specific; no finer-grained distinctions between roles are made.

PropBank annotation is closely tied to syntax. FrameNet (Baker et al., 1998) takes a different stance and focuses on accurate and detailed representation of event semantics. Verbs (as well as lexemes from other categories) are grouped into *frames* so that members of the same frame share a set of fine-grained frame-specific semantic roles (e.g. Impactee, Force, Buyer, Goods).

Both PropBank and FrameNet SRL operate on the verb sense/verb class generalization level. VerbNet (Schuler, 2006) groups verbs into Levin-inspired verb classes and defines sets of general, lexicon-level thematic roles and constructions for each class. It is the only SRL formalism that operates with a thematic role set. VerbNet role sets and verb class information are mapped to the PropBank corpus annotations via SemLink (Bonial et al., 2013).

2.3 Thematic roles in SRL

So far only few studies have considered VerbNet-level granularity in SRL and we are not aware of SRL systems specifically designed to exploit the thematic role generalizations. Zapirain et al. (2008) compare PropBank and VerbNet performance using a simple SRL system and conclude that PropBank labels generally perform better; however, they do not use any additional modeling possibilities offered by VerbNet's general, predicate-independent role set. Loper et al. (2007) show that replacing verb-specific PropBank roles A2-5 with the corresponding VerbNet roles improves the SRL performance. Merlo and van der Plas (2009) report a statistical analysis of PropBank and VerbNet annotations and conclude that while PropBank role inventory better correlates with syntax and is therefore easier to learn, VerbNet thematic roles are more informative and better generalize to new verb instances. Finally, a recent comparison on German data by Hartmann et al. (2017) positions VerbNet inventory above FrameNet and below PropBank in terms of complexity and generalization capabilities; however, the experiment is again based on the *mateplus* system (Roth and Woodsend, 2014) designed with PropBank generalization level in mind.

2.4 Semantic Proto-Roles

A related line of work is Semantic Proto Role Labeling (SPRL) (Reisinger et al., 2015; White et al., 2017) which, following Dowty (1991), discards the notion of atomic semantic role inventory and replaces it with Proto-Agent and Proto-Patient **property sets**. While our study utilizes traditional atomic role inventories, we see SPRL as a compatible parallel line of work and believe that additional benefits can be gained by combining the two views on syntax-semantics interface. In particular, Reisinger et al. (2015) investigate the alignment between Dowty-style role properties and VerbNet thematic roles and show that VerbNet Agents tend to bear Dowty's instigated, awareness and volitional properties, while Themes are more likely to change posession, change state, etc.

2.5 Thematic hierarchies

Numerous THs have been proposed in the linguistic literature, e.g. Agent $\prec$ Instrument $\prec$ Theme (Fillmore, 1968); see (Levin and Rappaport Hovav, 2005) for an overview. These hierarchies are rarely applicable for NLP since they originate from different theoretical backgrounds and are usually focused on a narrow set of linguistic phenomena (e.g. subject selection), aiming to provide a cross-linguistically valid hierarchy based on a set of manually constructed examples. In contrast, our approach is data-driven and aims to describe the general syntactic behavior of thematic roles. While an optimal TH that would successfully describe semantic roles' behavior across languages might not exist (and would imply the existence of a universal role inventory and grammar), our evidence suggests that this concept is at least partially applicable.

To the best of our knowledge, there exists no prior work explicitly aiming at discovering thematic hierarchies in corpora. However, the hierarchy-related effects are reported in some studies. For example, White et al. (2017) observe on a reduced role set that VerbNet roles dispref the violations of thematic/syntactic hierarchy alignment. Sun et al. (2009) experiment on thematic rank prediction for PropBank A0 and A1, but extend their analysis neither to VerbNet thematic roles, nor to the PropBank A2-5.

2.6 Syntactic formalisms

Cross-lingual applicability has traditionally been a strong component in semantic role theory, and universality is one of the common desiderata for a thematic hierarchy. This, however, implies the existence of a universal syntactic prominence scale.

From the NLP perspective, the closest to universal syntactic representation for which automatic parsers are available is the Universal Dependencies (UD) representation. Universal Dependencies (Nivre et al., 2016) is a recent initiative aimed at creating a single dependency-based formalism suited for describing syntactic structure in a language-independent way. It encompasses freely available treebanks for more than 60 languages, and universal dependency parsing is an active research area (Zeman et al., 2017). Based on that, we make an effort to ground our study in UD syntax for English. Since neither gold UD annotations, nor a deterministic converter are available, for German we use the TIGER dependency syntax representation (Dipper et al., 2001).

3 Hierarchical Linking model

3.1 Model

We suggest a simple model to describe the interface between syntactic and thematic rankings. An SRL corpus can be seen as a collection of sentences with corresponding **predications**, where each predication has a **target** (e.g. verb) and a set of **arguments** labeled with semantic roles.

Let $a_1...a_n \in A$ be the set of arguments in the predication p; $r(a_i)$ be the role label af the argument a_i, and $d(a_i)$ be the path between the predicate and the argument in the dependency parse tree of the sentence. A **syntactic ranker** S provides a syntactic rank $s_i = S(d(a_i))$ for each argument a_i in A based on the path, and a **thematic ranker** T provides a thematic rank $t_i = T(r(a_i))$ based on the argument's role. For each pair of arguments (a_i, a_j) we expect their syntactic ranks to align with their thematic ranks, i.e.

$$\forall i \neq j : sign(t_i - t_j) = sign(s_i - s_j)$$

The model per se does not imply the existence of a global ranking and allows flexible ranker definition. It allows ties in both syntactic and thematic rankings.

We use accuracy to assess how well a given syntactic-semantic ranker pair reflects the actual argument ranks found in data. Given a set of test predications $p_1, p_2...p_k \in P$ with the argument sets $A^1, A^2...A^k$, we measure the correspondence between syntactic and semantic ranking over the argument pairs (a_i^k, a_j^k) via accuracy defined as

$$\frac{\#(sign(t_i^k - t_j^k) = sign(s_i^k - s_j^k))}{\#total\_pairs}$$

To avoid the majority class bias, we measure accuracy for each role pair and use macro-averaged accuracy over pairs as the final score. A straightforward alternative to our evaluation metric would be the Kendall rank correlation coefficient, which, based on our preliminary experiments, tends to overemphasize the performance on most frequent role pairs.

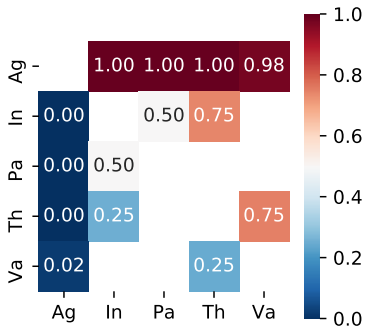

Figure 1: Preference matrix

4 Thematic Hierarchy Induction

This paper investigates several thematic ranking strategies. As a running example we use a small role set: `Agent` (`Ag`), `Patient` (`Pa`), `Instrument` (`In`), `Theme` (`Th`) and `Value` (`Va`). For now we assume the following syntactic hierarchy: *subj* $\prec$ *iobj* $\prec$ *nmod* $\prec$ *obj* $\prec$ *other*.

Local ranker The simplest way to model role ranking is to extract the average syntactic rank for each role based on the data, and then, given a test pair, assign ranks based on average syntactic rank.

role	Ag	Pa	In	Th	Va
$mean(s)$	1.01	2.58	1.72	3.95	3.74

Table 1: Mean syntactic rank per role (1-5)

Pairwise ranker Given that roles often strongly prefer a certain syntactic position (also see (White

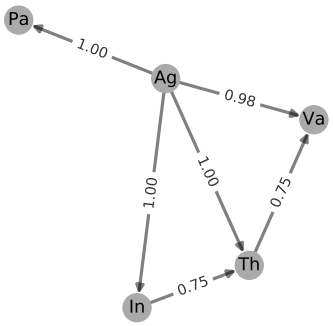

Figure 2: Preference graph

et al., 2016)), local ranking is a reasonable baseline strategy. However, it fails to account for the context dependency of thematic roles' syntactic realization. The next step is to construct a **pairwise preference matrix**: for each pair of roles encountered in training data we calculate the proportion of times role r_i receives a higher syntactic rank than role r_j. For our role set this results in the matrix shown on Fig. 1.

The preference matrix, for example, shows that `Agent` clearly dominates all the roles, `Instrument` ranks over `Theme`, and `Value` is below `Theme`.

Global ranker The pairwise ranking approach takes context into account. However, some role pairs only co-occur rarely. In such cases no pairwise ranking information is available to the model. Finding a global TH based on pairwise preferences is an example of a **rank aggregation** problem which can be solved via constrained ILP optimization on a **preference graph** (Conitzer et al., 2006). We represent the pairwise preference matrix as a graph $G = (v, e)$ where each vertex v represents a role, the edge weight is the preference strength measured as $\#(r_i \prec r_j)/\#(r_i, r_j)$. The edge direction is from higher- to lower-ranking role. If we assume a global ordering of the roles, we can induce the global ranking via transitivity relations. For example (Fig. 2), `Instrument` never appears with `Value` in our training data; however, by transitivity via `Theme` we can assume that `Instrument` ranks over `Value`.

Given the preference graph $G = (v, e)$, let w_{ij} be the weight of the edge between v_i and v_j. Let $x_{ij} \in 0, 1$ denote that we rank vertice v_i above v_j. The goal is then to maximize $\sum_{i,j} x_{ij} w_{ij}$ subject to two groups of constraints. First, we prohibit two nodes to rank above each other, but allow ties, by enforcing $\forall_{i,j} : x_{ij} + x_{ji} \leq 1$. Second, we enforce transitivity, i.e. if r_i is ranked above r_j, and r_j is ranked above r_k, then r_i must be ranked above r_k, formally $\forall_{i,j,k}, i \neq j \neq k : x_{ij} + x_{jk} - x_{ik} \leq 1$. We solve the ILP problem using the off-the-shelf *pulp* optimizer (Mitchell et al., 2011).

For our restricted example, optimization produces the following global hierarchy: `Ag` $\prec$ `In` $\prec$ `Th` $\prec$ `Va`/`Pa`. This hierarchy ranks `Instrument` above `Value` by transitivity, however, in case of `Patient` and `Value` no preference can be inferred from the graph, so they receive the same thematic rank.

5 Experiments

5.1 Datasets and Restrictions

For our experiments on English, we use SemLink (Bonial et al., 2013), a manually constructed resource that enriches PropBank's (Palmer et al., 2005) semantic role annotations with VerbNet's (Schuler, 2006) thematic role labels. We use the Universal Dependencies converter (Schuster and Manning, 2016) to transform original PropBank syntactic annotation to UD. PropBank semantic role annotation and the corresponding SemLink reference are constituents-based. However, UD is a dependency formalism, and we employ a number of heuristics to align original PropBank annotations with the CoNLL-2009 datasets (Hajič et al., 2009) to recover the head node positions. We employ additional transformations, filtering out the predications in which not all PropBank core roles got aligned to the VerbNet thematic roles.

For German, we use the recently introduced SR3de dataset (Mújdricza-Maydt et al., 2016; Hartmann et al., 2017) which explicitly provides VerbNet annotations on top of SALSA corpus (Burchardt et al., 2006). There exist no gold UD annotations for the SALSA corpus, and we use the SALSA's default TIGER syntactic formalism (Dipper et al., 2001) in our experiments.

Following previous work, we employ certain restrictions on our data. Since thematic roles in both VerbNet and SR3de are only defined for verbal predicates, we restrict the scope of our study to verbs. We only consider direct dependents of the verbs in active voice, and since having access to the full argument set is important to study con-

dataset	#sent	#tok	#pred	#arg
EN (PropBank→SemLink)				
train	16 603	446 641	21 276	44 333
test	1 031	27 751	1 336	2 761
dev	550	15 157	684	1422
DE (SR3de VerbNet)				
train	898	20 277	905	1 992
test	240	4 738	245	532
dev	117	2 429	119	266

Table 2: Dataset statistics

text dependency, we only consider the predications where all arguments are direct dependents of the verb in the UD dependency tree. Since we are interested in relative ranking, only predications that contain more than one semantic argument are considered in the study.

Dataset statistics for English and German (after filtering) are summarized in Table 2. In all experiments we induce a TH and related statistics from the training data and evaluate it on the test data, using the split from the CoNLL SRL shared tasks.

5.2 Syntactic ranker

For simplicity in this paper we only experiment with two syntactic rankers per language. A common syntactic prominence scale assumed in linguistic literature is *subject* $\prec$ *object* $\prec$ *indirect object* $\prec$ *oblique*. This scale has to be adapted to the UD and TIGER labeling schemes. For each language we evaluate two syntactic rankings: one that positions *objects* above *indirect objects* and *obliques*, and one that positions *objects* below.

For English, we rank the UD syntactic relations as follows (**SE1**): *nsubj / csubj* $\prec$ *iobj* $\prec$ *nmod* $\prec$ *ccomp / dobj* $\prec$ *other*; where *nmod* corresponds to oblique and *other* is used for any other syntactic relation. An alternative ranking positions *dobj* directly after the subject (**SE2**): *nsubj / csubj* $\prec$ *ccomp / dobj* $\prec$ *iobj* $\prec$ *nmod* $\prec$ *other*.

For German, the following ranking of TIGER syntactic relations is employed (**SD1**): *SB* $\prec$ *DA* $\prec$ *OP / MO / OG/ OC* $\prec$ *OA / OA2 / CVC* $\prec$ *other*; where *SB* is the subject, *DA* is dative object, *OP / MO / OG / OC* correspond to oblique relations, and *OA / OA2 / CVC* to direct object relations (see (Dipper et al., 2001) for detailed description). Similarly, we evaluate the performance of the ranking that positions the direct object after the subject (**SD2**): *SB* $\prec$ *OA / OA2 / CVC* $\prec$ *DA* $\prec$ *OP / MO / OG / OC* $\prec$ *other*.

	synt	glob	pair	loc	RND	UB
EN	SE1	.869	.887	.867	.509	.927
EN	SE2	**.930**	**.929**	**.913**	.500	**.932**
DE	SD1	.655	.726	.637	.471	.818
DE	SD2	**.790**	**.820**	**.820**	.456	**.920**

Table 3: Thematic ranker evaluation, incl. random ranker (RND) and upper bound (UB); bold - best result over syntactic rankers, underlined - best result over thematic rankers

5.3 Bounds

We construct the **upper bound** for the hierarchy induction by evaluating a global ranker trained on the test dataset. The upper bound reflects the data properties, as well as the maximal alignment accuracy that can be achieved with the selected syntactic ranker. The **lower bound** is constructed by evaluating 100 random thematic rankers which rank roles according to a random (but consistent) hierarchy, and averaging the result.

5.4 Data utilization setup

To evaluate how effective the proposed rankers use the training data, we conduct a series of experiments with reduced dataset sizes using the following protocol. The training dataset is shuffled and split into $n = 100$ slices. A ranker is consecutively trained on the first $m \in 1..n$ slices and evaluated against the full test dataset. The procedure is repeated $k = 100$ times to eliminate the effect of data order, and the results per slice are averaged.

6 Results

6.1 General Accuracy and Syntactic Ranker

To get an overall impression of the ranking quality, we first compare the performance of thematic rankers with respect to syntactic rankers and available datasets. The results of this comparison are summarized in Table 3 and show that syntactic rankers positioning the object second in the hierarchy (SE2 and SD2) lead to better alignment on both datasets and have a higher upper bound. We report the results on these rankers for the rest of the paper.

For English the global hierarchy-based ranker approaches the upper bound, closely followed by the pairwise ranker. The accuracy on German data is lower and the pairwise and local rankers outperform the global hierarchy-based ranker. We revisit this observation in 6.5.

EN	Agent ≺ Cause/Instrument/Experiencer ≺ Pivot ≺ Theme ≺ Patient ≺ Material/Source/Asset ≺ Product ≺ Recipient/Beneficiary/Destination/Location ≺ Value/Stimulus/Topic/Result/Predicate/Goal/InitialLocation/Attribute/Extent
DE	Agent ≺ Experiencer ≺ Stimulus/Pivot ≺ Cause ≺ Theme ≺ Patient ≺ Topic ≺ Instrument ≺ Beneficiary/InitialLocation ≺ Result ≺ Product/Goal ≺ Destination/Attribute ≺ Recipient ≺ Value/Time/CoAgent/Locus/Manner/Source/Trajectory/Location/Duration/Path/Extent

Table 4: Induced hierarchies

	EN-test	DE-test
UB	.932	.920
EN-train	.930	.787
DE-train	.852	.790
RND	.500	.456

Table 5: Cross-lingual evaluation, global ranker

Role pair	score	#(train)
Recipient - Topic	0.35	338
Source - Theme	0.46	246
Location - Theme	0.53	400
Material - Product	0.67	29
Result - Theme	0.67	30
Experiencer - Stimulus	0.74	922
Destination - Theme	0.86	401
Instrument - Theme	0.88	110
Recipient - Theme	0.89	419
Attribute - Experiencer	0.90	166

Table 6: Global ranker accuracy, English

6.2 Qualitative analysis

The result of hierarchy induction is a global ranking of thematic roles. Table 4 shows full rankings extracted for English and German data. While some correspondence to the hierarchies proposed in literature is evident (e.g. for English `Agent` ≺ `Instrument` ≺ `Theme`, similar to (Fillmore, 1968)), a direct comparison is impossible due to the differences in role definitions and underlying syntactic formalisms. Notice the high number of ties: some roles never co-occur (either by chance or by design) or occur on the same syntactic rank (e.g. *oblique*) so there is no evidence for preference even if we enforce transitivity.

6.3 Cross-lingual hierarchy induction

The induced hierarchies for English and German bear certain similarities, which raises the question on cross-lingual applicability of the hierarchies. This analysis is only possible because the VerbNet and SR3de role inventories are mostly compatible with few exceptions (Mújdricza-Maydt et al., 2016). Table 5 contrasts the performance of THs induced from English and German training data, and evaluated on German and English test data respectively. While the cross-lingual performance is expectedly lower than the monolingual performance, it outperforms the random baseline by a large margin, suggesting the potential for cross-lingual hierarchy induction.

6.4 Data utilization

One can assume that constructing a global hierarchy should require less training data due to the effective utilisation of transitivity. We evaluate this assumption empirically. Fig. 3 reports the performance of rankers with access to different amounts of training data for English and German. The results on English data show that global hierarchy-based ranker effectively utilizes the training data and can be trained using just fractions of the original training dataset.

The accuracy measurements on German are less conclusive: the local ranker generally performs best and learns fastest. We attribute this to the fact that filtered SR3de is an order of magnitude smaller than the PropBank/SemLink dataset. For pairwise and global rankers as many role pairs as possible should be observed at least once to establish the pairwise preference. This holds for PropBank/SemLink (all role pairs from test data seen at least once after observing 20% of the training data, on average), however, for filtered SR3de, even given the full training data, only 83% of role pairs from the test set have been seen at least once.

6.5 Error analysis

Our evaluation procedure allows detailed insights into the performance of the models. To illustrate, we extract the role pairs from English and German data with ranking accuracy below 1.0.

Table 6 lists the ranking inconsistencies produced by the global ranker for English. We can

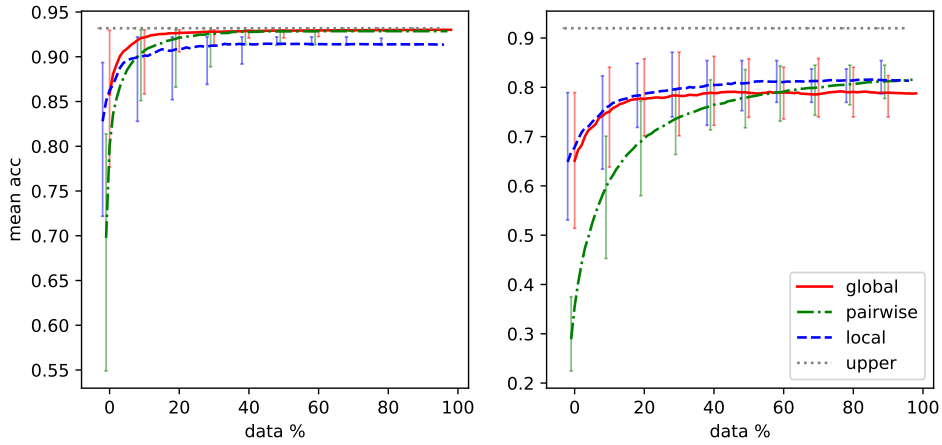

Figure 3: Data utilization for English (left) and German (right) along with max/min values

Role pair	score	#(train)
Attribute - Source	0.00	0
Beneficiary - Manner	0.00	0
Beneficiary - Value	0.00	1
Extent - Goal	0.00	2
Goal - Recipient	0.00	12
Instrument - Result	0.00	3
Locus - Topic	0.12	3
Recipient - Theme	0.40	26
Recipient - Topic	0.50	5
Pivot - Theme	0.67	57

Table 7: Global ranker accuracy, German

see that false ranking might be caused by the lack of training examples (e.g. `Material` vs. `Product`, `Theme` vs. `Result`). We also observe complications with positioning the `Theme` on the hierarchy. In many cases the misalignment is due to non-standard use of thematic roles, e.g. `Location` as subject in `wsj_2322:7` [*the* ***delay***$_{Loc}$ *resulted from* ***difficulties***$_{Th}$]. Another common reason for false alignments is the syntactic ranker. For example, in `wsj_2372:1` [*the* ***Senate***$_{Ag}$ *voted* ***87-7***$_{Res}$ *to* ***approve***$_{Th}$...] the `Result` is connected to the predicate via an *advmod* relation, and `Theme` is *xcomp*, both ranked equally (*other*) by our syntactic ranker.

Error analysis on the much smaller German dataset (Table 7) reveals the sparsity-related issues: most of the role pairs that tend to get misaligned do not, or only rarely appear in the training data, heavily influencing the score. As on English data, many misalignments are due to simplicity of the syntactic ranker.

7 Discussion

7.1 Importance of the syntactic ranker

The choice of syntactic ranking has a drastic effect on the resulting TH and the alignment quality, even if only direct syntactic dependents and a limited set of relations are taken into account. Realistically there might exist an arbitrary set of paths connecting arguments to predicates. UD as syntactic formalism is also subject to rapid change. Inducing a **joint syntactic and thematic hierarchy** that maximizes the overall alignment quality is a crucial direction for future work with potential benefits for SRL and syntactic parsing. Although we show that THs can be induced with an arbitrary dependency formalism, a **cross-lingual UD-based study** would be another extension to our work.

7.2 SRL integration

To utilize and evaluate the potential of thematic hierarchies for role interaction modeling, **SRL integration** is necessary. This, however, is not a trivial task: the absolute majority of semantic role labeling systems are designed with PropBank or FrameNet SRL formalism in mind and are not tailored to general VerbNet-style semantic roles and verb class-level disambiguation. A dedicated VerbNet SRL system would enable this assessment, and applying THs to such a system is an important future work direction.

7.3 Robustness to parsing errors

This paper focuses on TH induction using predefined syntactic annotation: a corpus annotated with semantic roles without an underlying syntactic layer is a rare occurence. However, for prac-

tical applications and for the cases when an SRL corpus is provided without syntactic annotations, it would be important to evaluate how effectively THs can be induced given parsing errors in training *and* in test data.

7.4 Data selection

We have demonstrated that THs can be induced from small portions of training data. The large discrepancy in the scores on the first data slices seen on Fig. 3 suggests that some data instances are more informative for TH induction. This raises the question whether it is possible to automatically **select useful training instances**, supported by the evidence from previous work in SRL (Peterson et al., 2014). One obvious strategy would be to make sure that the hierarchy inducer is presented as many role pairs as early as possible. Approximating this objective in an unsupervised way would reduce the amount of data needed to induce a high-quality thematic hierarchy.

7.5 The need for a global hierarchy

Our results regarding the **necessity of a global hierarchy** which ranks *all* the roles are inconclusive. While global ranking reaches the best quality for English, on the German data pairwise and local ranking approaches perform best. Although we attribute the latter to sparsity, more German data would be needed to evaluate this hypothesis. In particular, this can be achieved by relaxing some of the constraints we impose on the data.

8 Conclusion

This paper has presented an empirical framework for thematic hierarchy induction and evaluation. We have suggested several syntactic and thematic ranking strategies and a method to induce global thematic hierarchies from corpus data. Analysis on English and German data shows that hierarchy induction is feasible, data-efficient and has potential for cross-lingual applications. Promising directions for future work include joint modeling of syntactic and thematic ranking, selecting informative training instances and evaluating the utility of global hierarchies on extended language material.

Acknowledgements

This work has been supported by the German Research Foundation as part of the Research Training Group AIPHES (grant No. GRK 1994/1), QA-EduInf project (grant GU 798/18-1 and grant RI 803/12-1) and FAZIT Stiftung.

References

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*, volume 1, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006. The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 969–974. European Language Resources Association (ELRA).

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, FAM-LbR '10, pages 52–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vincent Conitzer, Andrew Davenport, and Jayant Kalagnanam. 2006. Improved bounds for computing kemeny rankings. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 620–626. AAAI Press.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 948–956, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefanie Dipper, Thorsten Brants, Wolfgang Lezius, Oliver Plaehn, and George Smith. 2001. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pages 24–41.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 76(3):474–496.

Charles J. Fillmore. 1968. The Case for Case. In Emmon Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, New York.

Jeffrey S. Gruber. 1965. *Studies in Lexical Relations*. Ph.D. thesis, MIT, Cambridge, MA.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 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*, CoNLL '09, pages 1–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Silvana Hartmann, Éva Mújdricza-Maydt, Ilia Kuznetsov, Iryna Gurevych, and Anette Frank. 2017. Assessing SRL frameworks with automatic training data expansion. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 115–121. Association for Computational Linguistics.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483. Association for Computational Linguistics.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Beth Levin and Malka Rappaport Hovav. 2005. *Argument Realization*. Research Surveys in Linguistics. Cambridge University Press.

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

Edward Loper, Szu ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1507–1516, Copenhagen, Denmark. Association for Computational Linguistics.

Paola Merlo and Lonneke van der Plas. 2009. Abstraction and generalisation in semantic role labels: Propbank, verbnet or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 288–296. Association for Computational Linguistics.

Stuart Mitchell, Michael OSullivan, and Iain Dunning. 2011. Pulp: a linear programming toolkit for python.

Éva Mújdricza-Maydt, Silvana Hartmann, Iryna Gurevych, and Anette Frank. 2016. Combining semantic annotation of word sense & semantic roles: A novel annotation scheme for VerbNet roles on German language data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*. European Language Resources Association (ELRA).

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

Daniel Peterson, Martha Palmer, and Shumin Wu. 2014. Focusing annotation for semantic role labeling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Michael Roth and Kristian Woodsend. 2014. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 407–413. Association for Computational Linguistics.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21, Prague, Czech Republic. Association for Computational Linguistics.

Weiwei Sun, Zhifang Sui, and Meng Wang. 2009. Prediction of thematic rank for structured semantic role labeling. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 253–256. Association for Computational Linguistics.

Aaron Steven White, Kyle Rawlins, and Benjamin Van Durme. 2017. The Semantic Proto-Role Linking Model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 92–98. Association for Computational Linguistics.

Aaron Steven White, Drew Reisinger, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Computational linking theory. *arXiv*, abs/1610.02544.

Benat Zapirain, Eneko Agirre, and Lluìs Marquez. 2008. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, June, pages 550–558. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajič jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. Association for Computational Linguistics.