Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5188–5197
Marseille, 20-25 June 2022

Thematic Fit Bits: Annotation Quality and Quantity Interplay for Event Participant Representation

Yuval Marton[1], **Asad Sayeed**[2]
[1]University of Washington, WA, USA
[2]Dept. of Philosophy, Linguistics, and Theory of Science, University of Gothenburg, Sweden
ymarton@uw.edu, asad.sayeed@gu.se

Abstract

Modeling thematic fit (a verb–argument compositional semantics task) currently requires a very large burden of labeled data. We take a linguistically machine-annotated large corpus and replace corpus layers with output from higher-quality, more modern taggers. We compare the old and new corpus versions' impact on a verb–argument fit modeling task, using a high-performing neural approach. We discover that higher annotation quality dramatically reduces our data requirement while demonstrating better supervised predicate-argument classification. But in applying the model to psycholinguistic tasks outside the training objective, we see clear gains at scale, but only in one of two thematic fit estimation tasks, and no clear gains on the other. We also see that quality improves with training size, but perhaps plateauing or even declining in one task. Last, we tested the effect of role set size. All this suggests that the quality/quantity interplay is not all you need. We replicate previous studies while modifying certain role representation details and set a new state-of-the-art in event modeling, using a fraction of the data. We make the new corpus version public.

Keywords: SRL, thematic fit, psycholinguistics

1. Introduction

Is more data always more effective than better annotation? Is it always cheaper just to obtain and use data with mid-quality annotation than improve annotation quality over a smaller dataset? Traditionally, to researchers grounded in linguistics, it seemed obvious that higher quality and richer annotation should be better. But with the advent of "Big Data", the common wisdom seem to have shifted toward more data; Deep Learning continued this trend (see examples in Section 1.1).

We re-examine these questions using two types of natural language processing (NLP) tasks: **(1)** supervised thematic role prediction (given a predicate and an argument's word span; here only its syntactic head word) and word prediction (given a predicate and a role); and **(2)** psycholinguistic tasks outside the explicit training objective: rating the thematic fit between a verb and its potential arguments. These tasks have a large body of work in computational linguistics (see section 1.2).

We examine the trade-offs in training models designed to accomplish these tasks through modeling events and their participants in a large corpus (task (1) above). The trade-off we focus on is using more data with mediocre linguistic annotations versus little data with higher-quality annotations. For the former, we replicate a PropBank-based model of Hong et al. (2018) using increasing subsets of their training data, a large corpus with machine-predicted annotations of mediocre quality. For the latter, we replace some annotation layers with equivalent layers generated by higher quality linguistic tools. Our model implementation differs from Hong et al. (2018) in how missing role information and unknown (out-of-vocabulary) roles are represented. Our replicated baseline is stronger than theirs. We also test whether training on the higher quality data keeps yielding better models as training set size increases. Last, we also look at the trade-offs in training models over increasingly richer, more fine-grained, but potentially sparser semantic role annotation.

1.1. The conundrum of data

Our first goal is to revisit a widely held assumption in the NLP community: mediocre machine-predicted annotation yields at scale better (or equivalent) models than high quality annotations (manual or state-of-the-art machine-predicted) whose scale is much smaller, due to compute, cost, and time constraints. McClosky et al. (2006) and Foster et al. (2007), *inter alia*, use self-training[1] to improve syntactic parsing – as an alternative to manually annotating more data – in same or different domain/genre. Petrov et al. (2010) show that using 100k machine-predicted constituency parses to train a new dependency parser contributed the equivalent of 2k manually annotated parses. Manually annotating is slower and more expensive but better by definition.

However, despite growing amounts of annotated and unannotated textual resources, a number of tasks with traditional linguistic levels of representation remain a challenge, with unsatisfactory performance in various research areas and applications: artificial intelligence (AI), machine reading / knowledge graph population, chatbots and natural language understanding (NLU), as

[1]Augmenting few manual annotations with ones predicted by a prior version of the same parser over a large text.

well as computational psycholinguistic modeling and computational linguistics.
How do quality and quantity affect our above-mentioned semantic and psycholinguistic tasks?

1.2. Semantic modeling

Our second goal is to explore ways to improve semantic modeling and the representations used for this modeling. We use *semantic modeling* to refer to tasks at the intersection of NLP and psycholinguistics that have to do with representing and processing generalized event knowledge (Pustejovsky, 1991; Zarcone and Padó, 2011) . The underlying tasks include:
Semantic role labeling (SRL) is the task of annotating text according to semantic frames and their roles as defined in frameworks such as FrameNet, VerbNet, or PropBank (Baker et al., 1998; Schuler and Palmer, 2005; Palmer et al., 2005). For example, given 'I cut the cake with...' (1) 'marzipan' or (2) 'a knife', the role *Instrument/A3-MNR* is normally desired for (2) but not for (1).
Role prediction: a simplified SRL task we use here. Instead of a sentence, the input is just a verb v and a noun n; optionally additional nouns and their thematic relation (role) with v; the expected output is the most likely thematic role of n with v in the same event.[2]
Slot / Role filling:[3] given a predicate (typically a verb) and a thematic role, what word or phrase would be most appropriate for that role? This task can be viewed as the complement of SRL (given the word or phrase, what is its role?).
Thematic fit: Given a predicate and a role (say, 'cut' + *Instrument*), how well would a speaker of a given language (here, English) find 'knife' or 'spoon' fitting to the given role? And by extension: given a subset of a predicate and arguments (optionally also modifiers), can we predict the typicality level of the most recently added member to the rest of the given subset? This is often an abstraction of sentence comprehension (in humans): our *thematic fit* estimation changes as we hear more of the uttered sentence (Amsel et al., 2015).
Indirect thematic fit estimation learning from SRL annotations has shown promising results (Hong et al., 2018; Tilk et al., 2016; Santus et al., 2017). Following Hong et al. (2018) and others, we consider only the arguments' syntactic heads together with their semantic roles.[4] Given a new role, predict the fitness level of all known words and use the score of the given filler in the full score distribution as a fitness rating. (We also look at the complement: given a new word, predict the fitness of each possible role). These predictions are scored relative to human judgments (see Section 2).

1.3. Contributions

Exploring the data requirements of modeling human semantic representations allows us to revisit the question of the inherent difficulty of semantic tasks: does semantic processing simply require more annotated data to achieve high quality, or has it not been represented in a way conducive for computers to learn adequately?

We look into how annotation quality and quantity (both the number of semantic frames and the number of sentences) affect learning. We also explore annotation granularity, taking thematic role set granularity (number of roles in model) as a test case. We often see that modelers focus on only the two most frequent PropBank semantic arguments (Arg0 and Arg1) and ignore the rest or lump the rest together under a catch-all tag. Similarly, they focus on only few modifiers (e.g., Tilk et al. (2016) and Hong et al. (2018) use 2 core roles and 3 modifiers). We therefore trained models with increasing numbers of thematic role types (the predicate, its core arguments, and often-optional, non-core arguments or modifiers), using the better taggers, parsers, and labelers, and observed changes in prediction quality.

To summarize our main contributions, we

1. test how quality of annotation affects supervised role/word prediction, as training set size increases. We show in small to large sizes that the mediocre annotation method is *not* as useful as better quality annotation.
2. test how quality of annotation affects thematic fit estimation as an application of our models that is not part of the training objective. We show that quality increases with training size but surprisingly, variance is high even in larger sizes, leaving no clear winner annotation.
3. claim that the high variance of the (indirectly optimized for) thematic fit estimation makes it more difficult to interpret conclusions from previous studies that did not report it.
4. show new state-of-the-art results on role and word prediction, as well as thematic fit estimate correlations.
5. tease apart effects of quality and quantity; tease apart the number of training sentences from the number of training frames (the semantic frames annotated in these sentences).
6. test how annotation granularity (role set size) affects thematic fit (and role/word prediction)
7. provide a new, open-data, large lexical semantic (and syntactic) resource in English, revising and expanding the previously published RW-Eng (Sayeed et al., 2018).

[2] A similar task – only without the input nouns – is the **(role) selectional preference** of the verb: what roles are more likely with this verb? E.g., 'cut': *Agent/Arg0, Patient/Arg1, Instrument/Arg3-MNR*.

[3] We interchangeably refer to it here as **word prediction**.

[4] We also follow them in using the simplified ProbBank roles (*A0, A1, ...*), unlike much of the thematic fit literature that uses *Agent, Patient,* etc. (Dowty, 1991).

2. Background and related work

Thematic fit norms take the form of averaged human-rated plausibility scores for a verb, a noun, and the noun's thematic role. For example, we ask human raters: how well does "sword" fit as an instrument with the verb "cut"? Thematic fit norms are a subset of semantic feature/property fit norms that pertain to verb-argument relations. Exploring thematic fit allows for exploring the structure of the human lexicon and for exploring generalizations about affordances and the relationship between world knowledge and compositional semantics.

McRae et al. (1998) collected an early set of thematic fit norms. Human raters were asked to use a 7-point Likert scale to judge the fit of particular nouns with particular verbs in given roles. These plausibility judgements focused mainly on *Agent-Patient* roles. Later Ferretti et al. (2001) provided norms for *Instrument* and *Location* roles.

Padó (2007) and Padó et al. (2009) sought to develop a probabilistic model of thematic fit. In the process, they collected additional *Agent-Patient* norms for a limited, balanced subset of verb-noun pairs chosen by frequency in the Penn Treebank (Marcus et al., 1993). Together, in addition to later efforts focusing on verb polysemy (Greenberg et al., 2015), these collected norms form an empirical basis for modeling human semantic expectations, albeit limited to roles that are relatively frequent and easily understood by raters.

Distributional modeling of thematic fit Early work in thematic fit modeling emphasized building partially or fully supervised corpus-based models (Padó and Lapata, 2007; Herdağdelen and Baroni, 2009). The question arises whether less task-specific, less supervised models can be used to model the semantic generalizations that would underpin a robust thematic fit model. Baroni and Lenci (2010) proposed the Distributional Memory (DM) approach, a very high-dimensional tensor space representation that memorizes the frequency of numerous syntactic relations between lexical items in a large corpus consisting of UkWaC (Ferraresi et al., 2008), the British National Corpus (Consortium and others, 2007, BNC), and Wikipedia.

Sayeed et al. (2016) applied the DM approach to relation features based in an early form of neural SRL tagger (Collobert et al., 2011, SENNA). This and Baroni and Lenci's syntax-based features were combined synergistically to produce thematic fit correlation scores superior to the result of each individually. The DM models were constructed without any reference to the evaluation data and can be considered unsupervised in that sense. However, their reliance on matrix multiplications made them difficult to extend to evaluating multiple roles simultaneously due to sparsity. They are also difficult to parameterize for finding optimal models.

Tilk et al. (2016) and Hong et al. (2018) worked to supplant DM approaches with neural networks. Their models train "event" embeddings with a preselected roleset, representing an entire semantic frame as input. Some of the role "slots" can be left empty, allowing for a variable number of arguments to be tested. Hong et al. (2018) applied a two-task training objective (limited SRL and role-filler noun prediction) to train NN models that not only performed well on thematic fit ratings, but also on several additional semantic tasks (e.g., event similarity and multiple-role compositionality).

Sayeed et al. (2016), Tilk et al. (2016), and Hong et al. (2018) all depend on the "Rollenwechsel-English", aka "RW-eng" corpus, hereafter `v1` (Sayeed et al., 2018), and use almost all of the corpus to train their models. Our work builds on the work of Hong et al. (2018), but differs in role set implementation, some hyperparameter settings, and minor other technical details. Our work also builds on `v1` and extends it with newer annotation layers. One of the things we test is whether the quantity of data used for these models is necessary to achieve those results, particularly on the thematic fit task.

3. Dataset

In order to explore the topics raised in Section 1, we used the above-mentioned large-scale `v1` corpus. It is annotated with a fast-but-outdated SRL tagger and syntactic parser. We added new annotation layers with higher quality, more modern taggers and parser (hereafter `v2`, "Our annotations"). We replicated baseline models on `v1` and trained new models on `v2`, as detailed in Section 4.

3.1. Text and `v1` Annotations

The `v1` corpus (Sayeed et al., 2018) consists of the SENNA-derived SRL output over 78M sentences from 2.3M documents. The documents come from the BNC and ukWaC. SENNA extracts multiple predicates per sentence and, for each predicate, it identifies spans of text representing noun phrases that fill PropBank roles for that predicate. For every document, sentence, and predicate in that sentence, `v1` contains XML-formatted information on the corresponding SENNA output. In particular, it uses a series of head-finding heuristics (Sayeed et al., 2016) to identify the syntactic heads of the role-filling spans—typically noun phrases, which can contain complex constituents such as subordinate clauses, but the SRL role spans could also cover only fractions of syntactic constituents, hence the need for a heuristic beyond only using a parser.

3.2. Our Annotations (`v2`)

NLP often contains "pipelines" of serial annotation processes, such as: tokenization and morphological analysis (including lemmatization or stemming), syntactic parsing, and a final processing such as machine translation, NLU (for chatbots), and sometimes SRL. As mentioned in Section 1, until about ten years ago, a rule of thumb often held: better quality of interme-

diate processing results in better quality at the end, although improvements are not linear, and small intermediate gains do not always translate to gains at the end of the pipeline.

Here we set to test out the new rule-of-thumb of the last decade for the case of thematic role prediction, slot filling (word prediction), and thematic fit estimation. We replaced annotation layers from older tools with annotations based on more recent tools (see below), introducing a non-negligible improvement in intermediate annotation quality.

The first step in doing so is to determine a consistent tokenization schema across all annotation layers, as some taggers either expect a certain schema or apply their own even if input text is largely tokenized. We iteratively modified our tokenization schema to reduce token count mismatch between the new layers from over 20% of sentences to less than 1%. Once we reached this low mismatch ratio, we marked and excluded the fewer mismatched cases from our experiments.

We added the following new annotation layers:

Lemmas by a newer morphological analyser: Morfette v0.4.4 (Chrupala et al., 2008; Chrupala, 2011), precompiled by Djamé Seddah,[5] who included a transformed xtag lexicon (Seddah et al., 2013). Training was done on the Penn Treebank (Marcus et al., 1993), using the Collins split.

Syntactic parses by a newer parser: spaCy 2.0.13 (Honnibal and Johnson, 2015; Honnibal and Montani, 2017).[6] We forced spaCy to use our own tokenization instead of its own.

Semantic frames by a newer SRL tagger: LSGN (He et al., 2018), an end-to-end BiLSTM-based SRL tagger using ElMo embeddings (Peters et al., 2018). It gets 86% F1 score on the CoNLL05 WSJ test set, compared to SENNA's 75%. For each semantic frame, we aligned the spaCy parses to each argument span in order to find the syntactic head of the span, using a similar heuristic as in `v1`. We only used the heads for modeling (see Section 4). We aligned each token in the argument span across all layers (surface word-form, Morfette lemma, spaCy lemma and entity (NER) tag, etc.)

3.3. Train / dev / test split

Of about 3500 files, few (less than 0.4%) were discarded due to processing issues, leaving us with 3490 files. Due to the fact that the corpus is comprised of more than one source, we assigned 16 files as the development set, and 16 as the test set, chosen uniformly[7]. The rest of the files were used as the full training set. Subsets of this training set were each chosen uniformly too, to emulate availability of smaller training sets. This split departs from Hong et al. (2018), which used the last 0.4% (14 files) as the test set, the immediately preceding 0.4% as the dev set, and the rest for training.

3.4. Additional test sets

We also tested our models on the above-mentioned thematic fit test sets, *without* optimizing on them:

Padó-all: A human-rated thematic fit score dataset collected with psycholinguistic motivations, created by Padó (2007) and containing 414 verb-noun-role triplets, where every two triplets differ only in the role, one of {*Arg0, Arg1, Arg2*}.

McRae-all: A similar dataset with human scores, created by McRae et al. (1998), containing 1,444 such triplets, grouped in pairs similarly, but the roles are only {*Arg0, Arg1*}, and the words are less frequent (a harder task).

4. Experiments

4.1. Baseline Model Configuration

For our baseline (`v1`-based models), we used a multi-task residual network (ResNet) model (He et al., 2016; Jégou et al., 2010). Our implementation is similar to the best reported model in Hong et al. (2018), called ResRofa-MT, for ease of comparison.[8] One task was **role prediction**, given a verb, a (typically noun) word, and zero or more ⟨role,word⟩ pairs. For example, given 'cut' and 'knife', predict *Instrument*, or in PropBank's roles: *A3-MNR*. A second task was **word prediction** (slot / role filling), given the word's role, the verb, and the zero or more ⟨role,word⟩ pairs. For example, given 'cut' and *A3-MNR*, predict 'knife'. For both tasks we could optionally provide more input, say, ⟨*A1*,'cake'⟩. Aside from having different prediction layer per task, the tasks shared the same neural network (and parameters). Apart from software engineering differences, the most notable difference in our implementation is having two separate labels for *missing role* and *unknown role* instead of one for both. The former is used to mark the absence of a certain role from the annotated frame instance. The latter is used as a catch-all for sparser roles not explicitly represented. The baseline role set was comprised of *PRD, Arg0, Arg1, ArgM-TMP, ArgM-LOC, ArgM-MNR*: the predicate, PropBank's arguments 0 and 1, the temporal, location, and manner modifiers (and *missing role* and *unknown role*).

For faster training time, we used a batch size of 1024 samples (unless otherwise specified) with the risk of too coarse updates. We kept a simple setting of 0.1 learning rate and no decay. We applied the same vocabulary pruning to the top 50k most frequent lemma forms as Hong et al. (2018) did.

4.2. `v1` vs `v2` experiments

We trained and evaluated models in the above configuration with increasing training set size (see Sec-

[5] We thank Djamé Seddah for making it available to us.

[6] https://spacy.io

[7] Dev files: [217, 435, 651, 868, 1085, 1302, 1519, 1736, 1953, 2170, 2387, 2604, 2821, 3038, 3255, 3472]. Test files: [218, 436, 652, 869, 1086, 1303, 1520, 1737, 1954, 2171, 2388, 2605, 2822, 3039, 3256, 3473].

[8] We thank Xudong Hong for his help.

Training sample (# trials)	Version	Role acc.	Word acc.	$\rho_{\text{Padó}}$ final	$\rho_{\text{Padó}}$ max	ρ_{McRae} final	ρ_{McRae} max
0.1% (3)	v1	.8857 ±.0009	.0435 ±.0001	.2760 ±.0331	.2760 ±.0331	.1924 ±.0110	.1968 ±.0124
	v2	.9102 ±.0063	.1029 ±.0007	.3149 ±.0308	.3257 ±.0412	.1934 ±.0044	.2065 ±.0057
1% (5)	v1	.9332 ±.0006	.0819 ±.0002	.5150 ±.0299	.5230 ±.0141	.3142 ±.0079	.3157 ±.0069
	v2	.9656 ±.0001	.1416 ±.0002	.4850 ±.0135	.4975 ±.0141	.3368 ±.0130	.3398 ±.0118
10% (3)	v1	.9419 ±.0017	.0941 ±.0005	.5166 ±.0345	.5368 ±.0020	.3996 ±.0206	.4126 ±.0091
	v2	.9715 ±.0010	.1541 ±.0045	.5229 ±.0227	.5623 ±.0227	.3935 ±.0192	.3981 ±.0223
20% (3)	v1	.9445 ±.0003	.0982 ±.0011	.5219 ±.0069	.5306 ±.0073	.4314 ±.0123	.4381 ±.0032
	v2	.9733 ±.0004	.1621 ±.0048	.5363 ±.0035	.5494 ±.0111	.4322 ±.0232	.4385 ±.0257

Table 1: v1 vs v2: model MTRFv4Res, train % out of 3490 text files, dev/test size = 16 text files each, batch=1024 unless specified. (# runs in parentheses). Role/word prediction accuracy (acc.); Spearman's rank correlation (ρ) for Padó-all and McRae-all. final:score of last saved model; max: maximal score in any epoch.

tion 3.3), measured in percentage of total number of available training sentences. For each training set size, we trained few models on v1 and same number of models on v2.

Table 1 shows that role prediction accuracy increases with training set size as expected and surpasses 90% at 1% of the training set for v1 and already at 0.1% for v2. At 10% of the training, it reaches mid-90s for v1 and high-90s for v2, with small further gains at 20%.[9] The advantage of v2 was kept throughout but decreased from over 5-6% (absolute) at 0.1% size to 3-4% at higher training set sizes. Word prediction accuracy followed similar trends, but v2−v1 gaps there remained about 6% in all set sizes.[10]

We also tested the models on two thematic fit tasks on which the models were not trained, using the additional test sets described in Section 3.4. Variance over multiple training runs per model was not reported in the relevant literature, so the high variance we see on both Padó-all[11] and McRae-all is a novel finding.[12] We report both the best- and last-epoch results for these tasks on the rightmost two columns of Table 1. We see v2 advantage on Padó-max 10% and 20% but not at lower sizes. Padó-final results are mixed, particularly at 10%-training, as are the results on McRae (both last and best). We see clear training size effects on McRae(best+last), but not on Padó (except in the smallest size).

The v1 results on the 10% and 20% subsets are our closest replication of Hong et al. (2018), modulo the above-mentioned role representation.

[9]The full dataset size was prohibitive for training given our computing resources.

[10]Preliminary experiments, having varying sized of dev and test sets, showed the same pattern: v2 advantage.

[11]In Padó-all evaluations, test set target *Arg2* was mapped to *unknown-role*, since it was not in the model's role set (McRae-all only has *Arg0,Arg1* targets).

[12]We trained 5 models for each version in small training sizes, and 3 for each in large sizes, due to resource limitations.

4.3. Roleset Granularity (on v2)

Model design includes decisions about features and their granularity. Generally, fine-grained features provide sharper, more accurate distributions of underlying phenomena, but their values, especially at the distribution "tail", suffer from higher sparsity, which may lead to difficulty in learning them well and hence lower performance overall. This tension between granularity and sparsity exists also for the case of role set granularity. Many studies have chosen the coarser side, using only few thematic roles. For example, Hong et al. (2018), which we take as our main baseline, use only 2 arguments and 3 modifiers, within the PropBank framework—in itself already a framework with a coarse role set (compared to VerbNet and FrameNet). We tested the effect of increasing role set granularity. In other words: is less (less coarse role set) more (higher quality)? We analyzed the distribution of the roles in the dev set (see Table 3) and expanded the role set one role at a time, according to their frequency (more frequent first). For training run time economy, all models in this subsection were trained on 1% of the v2 data.

The first semantic role to be added was *Arg2*. Note that now in Padó-all evaluations, *Arg2* is no longer mapped to *unknown-role*. See Table 2. It turns out that while role set prediction accuracy slightly dropped, word prediction accuracy actually improved by more than 1%. Same trend held also for Padó-all (about 3% gain in Spearman's correlation) and McRae-all (over 1%). This stands in contrast to preliminary experiments on v1.[13]

Next down the role list, we added *ArgM-MOD*. While we saw a slight gain in word prediction accuracy, we saw a drop in Padó-all and perhaps a slight drop in McRae-all. However, adding *ArgM-ADV* resulted in a drop in role prediction but gains in the thematic fit tasks. Adding *ArgM-DIS* resulted in some drop on the thematic fit tasks, while adding *ArgM-NEG* yielded relative gains in word prediction and Padó-all. Adding all

[13]Hong, personal communication.

Name	Role set	Role acc. dev/test	Word acc. dev/test	$\rho_{\text{Padó}}$ final/max	ρ_{McRae} final/max
2Args3Mods	baseline	**.9653 / .9656**	.1393 / .1414	.4765 /.4840	.3205 / .3240
3Args3Mods	+Arg2	.9595 / .9596	.1544 / .1563	**.5056 / .5150**	.3340 / .3340
3Args4Mods	+AM-MOD	.9606 / .9609	.1631 / .1661	.4663 / .4928	.3261 / .3373
3Args5Mods	+AM-ADV	.9513 / .9516	.1665 / .1691	.4838 / .5024	.3381 / .3407
3Args6Mods	+AM-DIS	.9503 / .9510	.1683 / .1712	.4742 / .4851	.3357 / .3370
3Args7Mods	+AM-NEG	.9506 / .9512	.1742 / .1768	.4808 / .4886	.3357 / .3385
all.args+mods	all-roles	.9450 / .9459	**.1783 / .1810**	.4833 / .5109	.3205 / .3209
3Args3Mods	+Arg2	.9595 / .9596	.1544 / .1563	.5056 / .5150	.3340 / .3340
4Args3Mods	+Arg3	.9580 / .9585	.1557 / .1582	.5007 / .5119	.3365 / .3394
5Args3Mods	+Arg4	.9574 / .9576	.1559 / .1583	.4901 / .5108	**.3473 / .3473**
6Args3Mods	+Arg5	.9577 / .9579	.1560 / .1582	.4925 / .5237	.3166 / .3182

Table 2: Increasing role set granularity (v2): model MTRFv4Res, train size = 1% of 3490 text files; dev/test size, batch size and metrics same as in Table 1. Top half: adding next role in descending role frequency. Lower half: adding only arguments (skipping modifiers). In each half, the roleset in each row is a superset of the previous row.

Count	Label
2,120,947	ARG1
1,234,063	PRD
1,090,751	ARG0
688,268	ARG2
380,294	ARGM-TMP
257,056	ARGM-MOD
227,040	ARGM-ADV
220,502	ARGM-MNR
194,532	ARGM-LOC
95,724	ARGM-DIS
87,036	ARGM-NEG
68,156	ARGM-PRP
39,780	ARGM-DIR
35,938	ARGM-ADJ
31,004	ARG3
27,850	ARGM-CAU
22,092	ARG4
18,254	ARGM-EXT
13,456	ARGM-PRD
9,108	ARGM-LVB
5,540	ARGM-GOL
3,826	ARGM-COM
3,460	ARGM-PNC
1,686	ARGM-REC
12	ARG5

Table 3: SRL label counts in dev set

data set	**previous (v1)**	**this (v2)**
training 10%	16,889,581	20,151,313
dev	766,333	915,473
test	767,325	919,365

Table 4: Number of frame annotations in v1 vs. v2

roles to the model (*all.args+mods*) yielded mixed results: a further small drop in role prediction (in fact, no model outperformed the baseline on this task), a pronounced gain in word prediction (highest result even compared to Table 1!), and lower scores on the thematic fit tasks compared to adding only *Arg2*.

Modifiers are often more freely optional, and are not considered core arguments of the semantic frame. We therefore also tested the effects of only adding core arguments to the baseline roleset (see bottom part of Table 2). Adding *Arg3* resulted in no much change in any task, compared to *+Arg2*. Adding *Arg4* yielded a gain on Padó-all last result (but not the maximal result). Adding *Arg5* resulted in a clear drop on the thematic fit tasks, which is expected: *Arg5* is very sparse, so its learnability is low, and roleset confusability is higher due to increased number of roles. However, we take differences on Padó-all and McRae-all with a grain of salt, given the above-mentioned high variability.

5. Discussion and Analysis

The advantage of v2-trained models over v1-trained models for role and word prediction in every training size undermines the NLP community's widely-held working assumption that mediocre is always preferable at scale. It supports our hypothesis that sometimes better annotations yield better results, even at scale, compared to baselines of reasonable mediocrity. While we do not know if this holds also in the limit, this finding is worth keeping in mind even with today's very large datasets.

To validate our claim that v2 annotations are much better than v1, we randomly sampled 8 sentences, and counted the difference in number of frames, number of roles/arguments, and number of wrong roles between the two datasets. v2 had a clear advantage over v1 in identifying frames and roles, with almost no cases in which v1 did better. This advantage held in both the number of cases and the number of sentences with offending cases (63–75%). Both v1 and v2 had few wrong roles, similar in number, with perhaps a slight advantage to v1 (one case).[14] Due to the small sample size and evaluation method, we take the findings above as mainly qualitative, but still strongly supporting our assumption of v2 advantage.

Was **quality** the only factor? Several aspects here: **(a)** better argument span and role prediction in LSGN in v2 compared to SENNA in v1, together with **(b)**

[14]Note that for verification speed, we only verified correctness of v2's *Arg0, Arg1*; the rest we assumed correct.

the greater *quantity* of predicted frames with LSGN, compounded by **(c)** better parsing quality of spaCy in `v2` compared to Malt in `v1`, and **(d)** Morfette's better lemma analysis.

As for **quantity**, it turns out it may have also played a role: we compared the number of semantic frame annotations in the 10% training subset, and found out `v1` has less than 84% of `v2`'s, for the same underlying sentences (Table 4). We call it **frame quantity** to tease it apart from **sentence quantity**: the number of (underlying) sentences used for training. Perhaps the better SRL and parser quality contributed both to the increase in number of frames as well as to the number of correctly extracted syntactic head words (one per argument, using better aligned parses).

Was the `v2` advantage mainly due to frame quantity? 1%-training `v2` outperforming 10%-training `v1` on role and word prediction, even though the former was trained on an eighth of the number of frames (and tenth of the sentences), suggests otherwise. This trend repeated with 0.1% `v2` outperforming 1% `v1` on word prediction. Higher quality annotations resulted in large *savings* in training sentence quantity for similar prediction quality.

How does our implementation fare compared to Hong et al. (2018)? Our `v2` maximal result on Padó-all (59.9% best single run, 56.2% averaged) outperforms their reported 53% with only 10% of their data. On McRae-all our maximal result (45.9% best single run, 43.9% averaged) outperforms their reported 42.5%, despite having less frequent pred-arg combinations. Our `v2` outperforms their reported 94.7% role accuracy, even at 1% of their training data, presumably due to our separating the *unknown role* from *missing role*.

Could we have reached even higher results? Our model setting is on the simpler side with only two tasks, one of which is role prediction with accuracy approaching 100% (therefore, after a few iterations, largely only the other task affects the learning). More complex models or additional tasks (and/or modern word embeddings) are likely to do even better on the word prediction and the thematic fit tasks. However our focus in this work was not on creating the best model, but on exploring the effects of annotation quality and quantity up to large scale.

Did the effort of creating `v2` pay off also for the **psycholinguistic task**? We see a clear, but rather small gain in Padó-all on the larger subsets (10%, 20%). In McRae-all the gain is actually on the smaller subsets, and goes away on the larger ones. Therefore, we conclude that for improving indirectly-supervised psycholinguistic tasks, the cost-effectiveness of this exercise is questionable, but it still suggests that progress can be made in resource-constrained environments through limited improvements in label accuracy.

As for **annotation granularity**, adding *Arg2* to the baseline's role set showed clear gains across the board (except perhaps in role prediction), which may seem surprising at first: **(a)** the role prediction task is now harder (larger role set) yet accuracy did not drop by much. This could be due to the increased thematic homogeneity in the catch-all role tag. **(b)** Word prediction (slot-filling) was not expected to be affected, since the embeddings are not relearned in our setting. But gains may show if many words tend to assume only certain roles (e.g., be Arg2-centric). **(c)** Role prediction in Padó-all should have also been harder, therefore showing lower correlation scores. But recall that baseline scores are not directly comparable here because all roles but Arg0 and Arg1 were mapped to Arg2 for the evaluation of Padó-all, a mapping which was no longer needed once we added Arg2 to the role set. **(d)** gain in McRae-all is surprising at first because McRae-all only has Arg0 and Arg1 targets, so adding Arg2 could only distract from these targets. But recall we have a catch-all role, and adding Arg2 to the role set made the catch-all role distribution more focused, and therefore presumably less prone to mix-ups with Arg0 and Arg1 – especially since McRae-all contains less frequent words, which makes them harder to learn.

Adding the next modifiers in order of descending frequency (top part of Table 2) yielded consistent monotonous gains in word prediction, with the *all-roles* model, trained on 1% of the training set, performing even better than larger subset models in Table 1. This seems to support finer-grained representation. However, curiously, adding only core arguments (lower half of Table 2) did not make a noticeable difference on this task. Adding the next modifiers after *Arg2* did not improve Padó-all, which is expected, since it only contains arguments 0-2. But the top result of *+Arg4* on McRae-all is surprising: this test only has *Arg0, Arg1* targets. We suspect this result is an outlier even given the high variance, but further investigation is in order.

Ethical considerations This work used two large corpora (the BNC and ukWaC); hence it is not practical to completely account for all the data in the corpus. The BNC is a curated corpus, but part of their transcribed conversations were recorded without prior consent of all recorded individuals. This is no longer an acceptable conduct in Great Britain and many other countries. Our annotated corpus (`v2`) clearly marks the source of each sentence, so those who wish to exclude BNC data can easily do so. Insofar as future work keeps that in mind, we believe there to be minimal scope for direct misuse of our results.

6. Conclusions and Future Work

We set out to test the NLP community's widely-held assumption that mediocre linguistic annotation at scale is as good as better annotation. We saw that models trained on better lemmas, syntactic parses, and SRL tags (our `v2`) did better than the baseline (`v1`) using older technology at all training set sizes, and even at scale – on both (directly supervised) role and word prediction. We also saw "training dataset savings"

potential: training on smaller sets with better annotations yielded sometimes better results than training on datasets with less advanced annotations that were 2 to 10 times larger in size (with McRae-all being a notable exception). To better understand that, we teased apart contributions of annotation quality and quantity, and their interplay. We further teased apart sentence quantity from frame quantity.

We saw a high variance in thematic fit estimation, to the point where in one task (Padó-all) `v2` advantages over `v1` only showed in larger training set sizes, while in another task (McRae-all) no advantages were seen. Given the small-to-no gains in these tasks relative to `v1`, the cost-effectiveness of re-annotating with better tools is questionable for the indirect supervision setting.

We saw all tasks benefited from increasing the training sentence set size, at least until 10% of our large corpus (except perhaps Padó-all beyond 1%). Future work should check if larger sizes can yield even better results. Even at 10% of the training, our `v2` model set a new record in indirectly supervised thematic fit estimation on Padó-all, and at 20% a new record also on McRae-all (both `v1`, `v2`).

We also saw that refining the semantic role set granularity helps in thematic fit tasks (and word prediction). On Padó-all, best results were achieved already by adding *Arg2*, but surprisingly on McRae-all by adding *Arg2-4*. Adding all roles yielded best results on word prediction. These are novel results.

Last, but not least, we introduced a new open-data annotated corpus.[15] We believe this new corpus will be useful to the NLP community beyond our reported experiments. Future work involves studies with existing and new annotation layers, e.g., combining `v1`+`v2` parses, `v1`+`v2` SRL tags, and replicating these experiments with various word embeddings and different network architectures. Future work should also explore different optimization objectives or additional tasks in the multi-task setting, since role prediction seems too easy (reaches accuracy in high 90s early on), while the word prediction objective may be too hard, although the latter can be ameliorated with better word embeddings and a loss function based on the vector distance between predicted and target words. We also plan to add new thematic fit tasks with multiple simultaneous role-fillers (Bicknell et al., 2010; Vassallo et al., 2018).

7. Acknowledgements

A. Sayeed's involvement in this research was funded by a Swedish Research Council (VR) grant (2014-39) for the Centre for Linguistic Theory and Studies in Probability (CLASP). We thank the reviewers for helping us make this paper clearer.

8. Bibliographical References

Amsel, B. D., DeLong, K. A., and Kutas, M. (2015). Close, but no garlic: Perceptuomotor and event knowledge activation during language comprehension. *Journal of Memory and Language*, 82:118–132.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, page 86–90, USA. Association for Computational Linguistics.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63(4):489–505.

Chrupala, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Chrupala, G. (2011). Efficient induction of probabilistic word classes with LDA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 363–372, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Consortium, B. N. C. et al. (2007). British national corpus version 3 (bnc xml edition). *Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved February*, 13:2012.

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

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Foster, J., Wagner, J., Seddah, D., and Van Genabith, J. (2007). Adapting WSJ-trained parsers to the british national corpus using in-domain self-training. In *Proceedings of the 10th International Conference on Parsing Technologies*, pages 33–35. Association for Computational Linguistics.

Greenberg, C., Demberg, V., and Sayeed, A. (2015). Verb polysemy and frequency effects in thematic fit modeling. In *Proceedings of the 6th Workshop on*

[15] Available at `http://yuvalmarton.com/RW-Eng`

Cognitive Modeling and Computational Linguistics, pages 48–57, Denver, Colorado, June. Association for Computational Linguistics.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

He, L., Lee, K., Levy, O., and Zettlemoyer, L. (2018). Jointly predicting predicates and arguments in neural semantic role labeling. *CoRR*, abs/1805.04787.

Herdağdelen, A. and Baroni, M. (2009). Bagpack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 33–40. Association for Computational Linguistics.

Hong, X., Sayeed, A., and Demberg, V. (2018). Learning distributed event representations with a multi-task approach. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 11–21, New Orleans, Louisiana, June. Association for Computational Linguistics.

Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September. Association for Computational Linguistics.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jégou, H., Douze, M., and Schmid, C. (2010). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June.

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159. Association for Computational Linguistics.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Padó, U., Crocker, M. W., and Keller, F. (2009). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Padó, U. (2007). *The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Human Sentence Processing*. Ph.D. thesis, Saarland University.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics*.

Petrov, S., Chang, P.-C., Ringgaard, M., and Alshawi, H. (2010). Uptraining for accurate deterministic question parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713, Cambridge, MA, October. Association for Computational Linguistics.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, 17(4):409–441.

Santus, E., Chersoni, E., Lenci, A., and Blache, P. (2017). Measuring thematic fit with distributional feature overlap. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 659–669, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sayeed, A., Greenberg, C., and Demberg, V. (2016). Thematic fit evaluation: an aspect of selectional preferences. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 99–105, Berlin, Germany, August. Association for Computational Linguistics.

Sayeed, A., Shkadzko, P., and Demberg, V. (2018). Rollenwechsel-English: a large-scale semantic role corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*.

Schuler, K. K. and Palmer, M. S. (2005). *Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, USA. AAI3179808.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J., Farkas, R., Foster, J., Goenaga, I., Gojenola, K., Goldberg, Y., et al. (2013). Overview of the SPMRL 2013 shared task: cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the 2013 SPMRL Workshop*. Association for Computational Linguistics.

Tilk, O., Demberg, V., Sayeed, A., Klakow, D., and Thater, S. (2016). Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 171–182, Austin, Texas, November. Association for Computational Linguistics.

Vassallo, P., Chersoni, E., Santus, E., Lenci, A., and Blache, P. (2018). Event knowledge in sentence processing: A new dataset for the evaluation of argu-

ment typicality. In *LREC 2018 Workshop on Linguistic and Neurocognitive Resources (LiNCR)*.

Zarcone, A. and Padó, S. (2011). Generalized event knowledge in logical metonymy resolution. In Laura A. Carlson, et al., editors, *Proceedings of the 33th Annual Meeting of the Cognitive Science Society, CogSci 2011, Boston, Massachusetts, USA, July 20-23, 2011*. cognitivesciencesociety.org.