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

Yuval Marton — University of Washington
Asad Sayeed — University of Gothenburg

Background

- With advent of Big-Data (and faster more powerful hardware), the received wisdom is that more (so-so labeled) data = better models.
- More "silver" labeled data $\geq$ small expert-annotated "gold" labeled data.
- More crowdsourced lay-person labels ¿= small expert-annotated labels.
- More labeled data by mediocre taggers ¿= less labeled data by SotA taggers.
- McClosky et al. (2006), Foster et al. (2007), Petrov et al. (2010), . . .

Research questions

Does "more data = better" / "better labeling insignificant at scale" apply to:

- Semantic role prediction? (SRL "lite": given a verb and arg, what's its role?).
- Role filling / word prediction? (given predicate and role).
- **Thematic fit estimation?**
- **Especially as a related task the model was not directly optimized for?**

Lexical Resource Improvements

- Original corpus: RW-Eng v1, Sayeed et al. (2018).
- Relabelling:

Original (v1)	**Replaced with (v2)**
NLTK/WordNet (Bird et al., 2009)	Morfette (Chrupała et al. 2008)
MaltParser (Nivre et al. 2006)	SpaCy (Honnibal and Montani, 2017)
SENNA (Collobert et al., 2011)	LSGN (He et al., 2018)

- Extensive work aligning of tokenization schemas over 78M sentences.

Quality vs. quantity

- Better span/role prediction in LSGN vs. SENNA.
- 20% more frame quantity with LSGN.
- Better parsing quality (spaCy vs. MaltParser) and lemmatization (Morphette).

Data quantity:

- Same number of sentences.
- **1% training `v2` outperformed 10% training `v1` with an eighth of the frames.**

Tasks

Task	**Input**	**Output**	**Comments**
Role-prediction ("SRL-lite")	Predicate, arg (head)	role	"child eat apple" $\rightarrow$ prob of Agent, Patient, . . .
Role/slot-filling (word prediction)	Predicate, role	arg head (lex item)	"child eat Patient" $\rightarrow$ prob of "apple", "cake", . . .
Thematic fit (Padó and McRae norms)	Predicate, argh (head), role	Score [0..1]	"child:Agent eat dog:Patient" $\rightarrow$low score for dog in frame+role. Few tests sets; no training data!

Results

				$\rho_{Padó}$		ρ_{McRae}	
Size	Ver	Role acc.	Word acc.	final	max	final	max
0.1%	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%	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%	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%	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

ρ = Spearman's ρ

Research answers

- **Training data savings**: improving annotation quality reduces data requirement up to 10-fold for role and word prediction.
- Models trained on better lemma identification, better parsing, better SRL tags did better than baseline at all most data sizes.
- **We are releasing a large resource with modern annotation: RW-English v2**