RUG-1-Pegasussers at SemEval-2022 Task 3

Data Generation Methods to Improve Recognizing Appropriate Taxonomic Word Relations

Frank van den Berg, Gijs Danoe, Esther Ploeger and Wessel Poelman

Supervised by Tommaso Caselli and Lukas Edman

RUG-1-Pegasussers at SemEval-2022 Task 3: Data Generation Methods to Improve Recognizing Appropriate Taxonomic Word Relations

Frank van den Berg*, Gijs Danoe*, Esther Ploeger*, Wessel Poelman*
Lukas Edman, Tommaso Caselli
Department of Information Science
University of Groningen
{f.l.van.den.berg, g.danoe, e.ploeger.1, w.g.poelman}@student.rug.nl
{j.l.edman, t.caselli}@rug.nl

Abstract

This paper describes our system created for the SemEval 2022 Task 3: Presupposed Taxonomies - Evaluating Neural-network Semantics. This task is focused on correctly recognizing taxonomic word relations in English, French and Italian. We developed various data generation techniques that expand the originally provided train set and show that all methods increase the performance of models trained on these expanded datasets. Our final system outperformed the baseline system from the task organizers by achieving an average macro F1 score of 79.6 on all languages, compared to the baseline's 67.4.

1 Introduction

In this paper, we describe our system and approach for the SemEval 2022 PreTENS (Presupposed Taxonomies: Evaluating Neural Network Semantics) shared task.[1] The aim of this task is to gain a better understanding of the ability of language models to recognize taxonomic relations between two words.

We focus on subtask 1, which is a binary classification task in which a system should predict whether a sentence is valid or not, depending on the taxonomic word relation in a given sentence. fine-tune a base English BERT (Devlin et al., 2019) model for the final classification task.

In our approach, we incorporate all three languages for this task: English, Italian and French. Instead of generating augmented training sets for each language and training separate models, we opted to train an English model and translate the Italian and French sentences to English, before predicting the validity labels. We chose this approach in part because several of our data generation methods were not available for French or Italian. We made use of Google Translate, as this is a widely used state-of-the-art general-domain translation system. Our model, trained on the expanded dataset, scored an average F1 score across all languages of 79.6, which is an improvement over the 67.4 baseline score. We found that the best data expansion technique was to combine multiple approaches, where the output of one method was the input for the next. Our ablation experiments show that our paraphrasing method improved scores the most. All code, data and other related files can be found in our GitHub repository.[2]

2 Task description

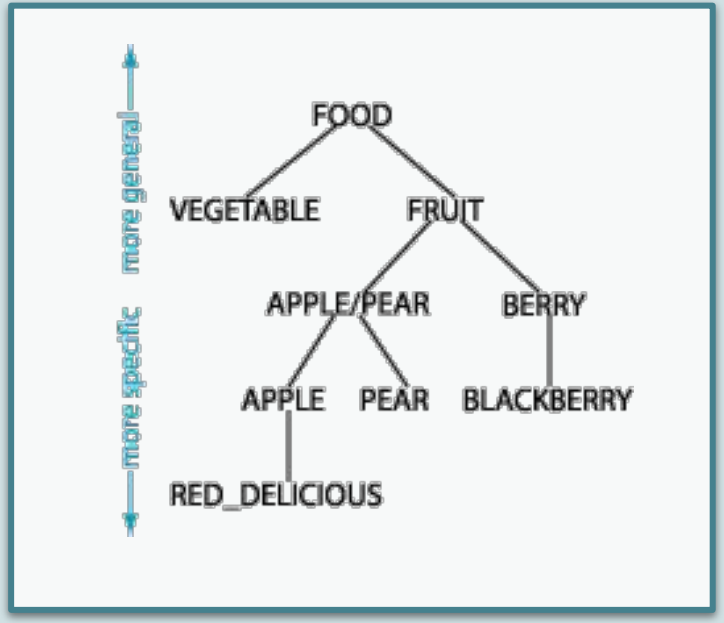
For the binary classification subtask, the challenge

OVERVIEW

- Task Description
- Method
- Results
- Analysis
- Discussion & Conclusion

Task description

- Recognizing taxonomic word relations
- Probing task



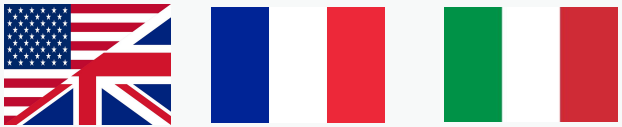
Task description

Binary classification

I like **trees** and, more specifically, **oaks.** → 1

I like **oaks** and, more specifically, **trees.** → 0

Languages



Method

Research Question

What are effective data generation approaches in order to improve a language model's ability to recognize appropriate taxonomic word relations?

Method

How to evaluate our experiments?

Official training data

- 5,837 samples
- 7 templates
- 14 noun categories
- 3 types of constraints

Method

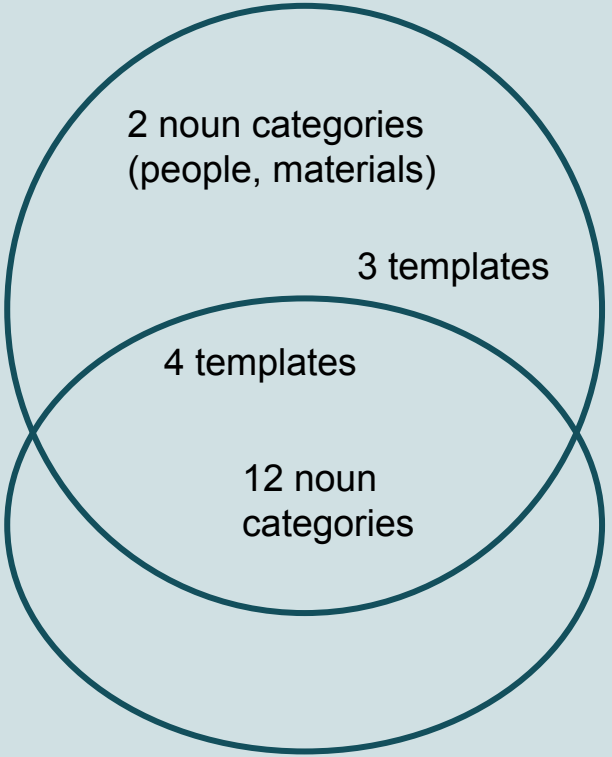
How to evaluate our experiments?

Official training data

- 5,837 samples
- 7 templates
- 14 noun categories
- 3 types of constraints

Development set
(3,101 samples)

Training set
(2,736 samples)



Method

Data generation

I like X more than Y

I prefer X over Y

(etc.)

Adding new templates

I like ham but not fish

I like **melon** but not **cheese**

Adding new words

I like pork, except **bacon**

I like **bacon**, except bacon strips

Hyponym of hyponyms

Method

Data generation

I like **animals**, except **pigs**
LABEL: 1

I like **pigs**, except **animals**
LABEL: 0

Inversion

I like beds, an interesting type of furniture

Beds are an interesting type of furniture that I like.

Paraphrasing

Method

	Sentences	F1
Individual methods		
Train	2,737	53.1
Train, hyponyms	4,957	55.5
Train, inverted	2,868	61.7
Train, new words	21,456	65.4
Train, templates (new words)	9,000	71.3
Train, templates (only original train set words)	9,000	73.1
Train, new words lemmatized*	21,456	73.1
Train, pegasus	19,484	77.6
Combined methods		
Train, hyponyms, templates	18,831	70.0
Train, new words, templates	21,456	74.7
Full pipeline combinations		
Templates, new words, inverted, pegasus	138,572	57.9
Templates, new words, hyponyms, pegasus, lemmatized*	211,354	59.1
Templates, new words, hyponyms, inverted	40,820	62.2
Templates, new words, hyponyms, inverted, pegasus	282,834	81.5
Templates, new words, hyponyms, inverted, pegasus, lemmatized*	282,796	83.6
Templates, hyponyms, inverted, pegasus	147,008	85.6
Templates, new words, hyponyms, pegasus	211,354	**88.1**

Results on 'development set'

Method

	Sentences	F1
Individual methods		
Train	2,737	53.1
Train, hyponyms	4,957	55.5
Train, inverted	2,868	61.7
Train, new words	21,456	65.4
Train, templates (new words)	9,000	71.3
Train, templates (only original train set words)	9,000	73.1
Train, new words lemmatized*	21,456	73.1
Train, pegasus	19,484	77.6
Combined methods		
Train, hyponyms, templates	18,831	70.0
Train, new words, templates	21,456	74.7
Full pipeline combinations		
Templates, new words, inverted, pegasus	138,572	57.9
Templates, new words, hyponyms, pegasus, lemmatized*	211,354	59.1
Templates, new words, hyponyms, inverted	40,820	62.2
Templates, new words, hyponyms, inverted, pegasus	282,834	81.5
Templates, new words, hyponyms, inverted, pegasus, lemmatized*	282,796	83.6
Templates, hyponyms, inverted, pegasus	147,008	85.6
Templates, new words, hyponyms, pegasus	211,354	**88.1**

Results on 'development set'

We discard inversion for our final system submission

Method: overview

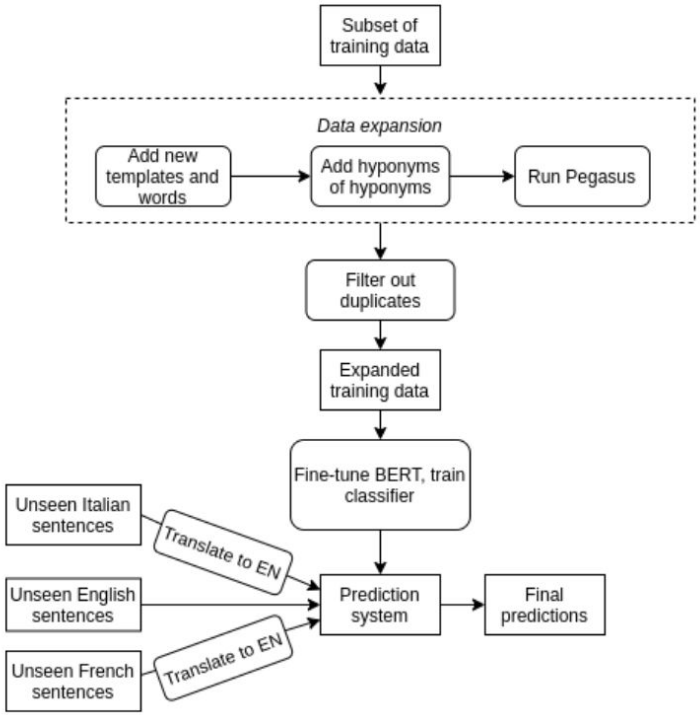


Figure 1: Overview of system and prediction pipeline

Results

Language	Precision	Recall	Macro F1
Italian	83.42	69.25	75.682
French	84.21	70.96	77.019
English	86.07	70.44	77.474

Results on official test set

Results

Sub-task 1 Global Ranking

Rank	username	Team Name	Members	Institution	Global Rank
1	wengsyx	LingJing	Yixuan Weng , Bin Li , Fei Xia	CASIA and Hunan University	94.4852
2	yingluLi	HW-TSC	Yinglu Li, Min Zhang, Xiaosong Qiao, Minghan Wang	2012 Labs, Huawei	92.7968
3	injySarhanUU	UU-TAX	Injy Sarhan, Pablo Romero, Marco Spruit	Utrecht University	91.5744
4	csecudsg				91.1167
5	piano				90.736
6	holdon	SPDB Innovation Lab	YueZhou, BoweiWei, JianyuLiu, Yangyang	Shanghai Pudong Development Bank	89.5786
7	bpc				89.0892
8	weijiyao				88.778
9	ddd7788				86.6832
10	cnxupupup				86.6762
11	robvanderg	MaChAmp	Rob van der Goot	IT University of Copenhagen	86.4218
12	aidenqiu				86.297
13	thanet.markchom	UoR-NCL	Thanet Markchom, Huizhi Liang, Jiaoyan Chen	University of Reading, Newcastle University, University of Oxford	80.3184
14	wpoelman	RUG-1-pegasussers	Esther Ploeger, Frank van den Berg, Gijs Danoe, Wessel Poelman	University of Groningen	79.5551
15	breaklikeafish	KaMiKla	Karl Vetter, Miriam Segiet, Klara Lennermann	Eberhard Karls Universität Tübingen	77.9852
16	huawei_zhangmin				71.7989
17	dbusca	RCLN	Davide Buscaldi	Université Sorbonne Paris Nord	70.5367
18	aridhasan	Organizers			67.394
19	borisdejong	Jan/Jasper/Boris	Jan Harms, Jasper Bultman, Boris de Jong	University of Groningen	27.255
20	folkertleistra				22.6408
21	RobertGroningen				19.9501

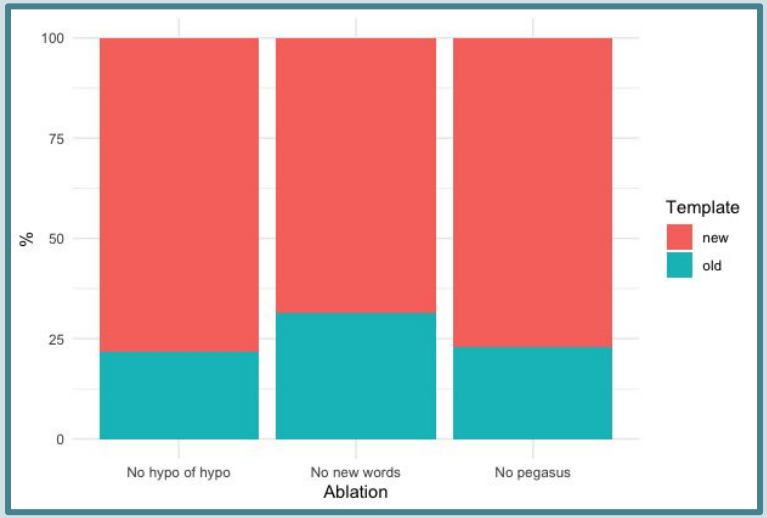
Analysis

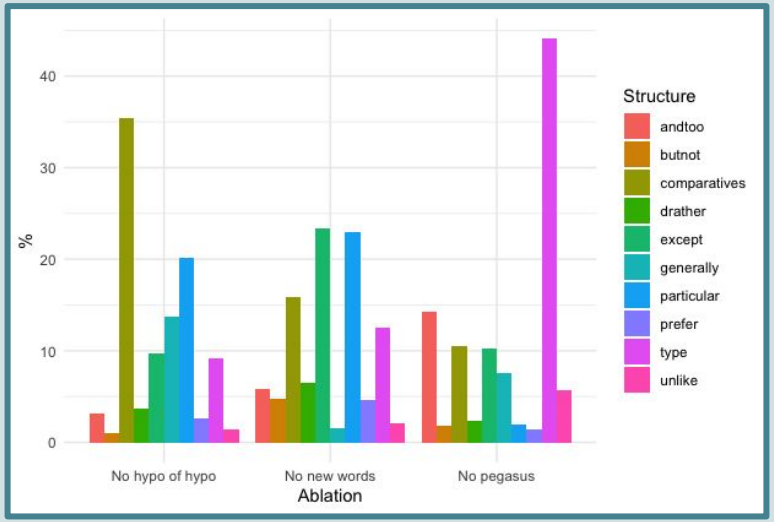
Ablation experiments

	Sentences	Acc	Pre	Rec	F1
Templates, new words, hyponyms, pegasus (used for final submission)	211,354	80.7	86.0	70.4	77.4
- pegasus	33,571	77.2	80.0	68.9	74.0
- new words	108,819	81.9	84.1	75.9	79.8
- hyponyms	116,234	79.0	79.9	74.1	76.9

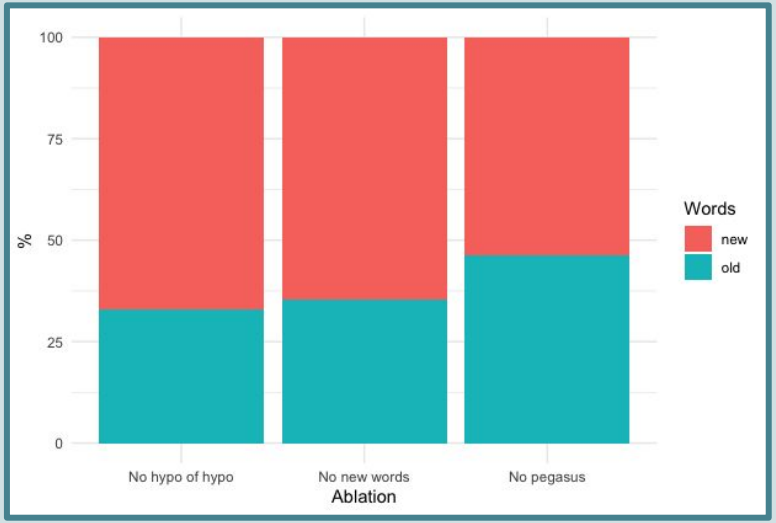
Ablation test with final expanded training dataset on official test set

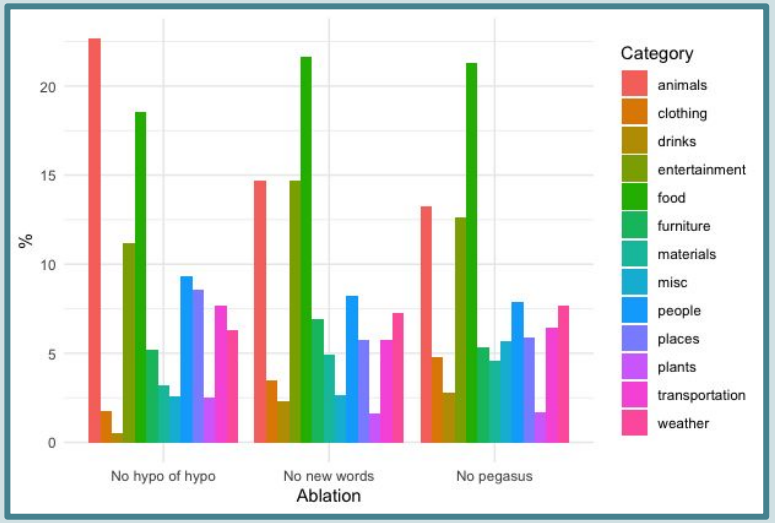
Analysis


100
75
50
25
0
%
No hypo of hypo
No new words
No pegasus
Ablation
Template
new
old


40
30
20
10
0
%
No hypo of hypo
No new words
No pegasus
Ablation
Structure
andtoo
butnot
comparatives
drather
except
generally
particular
prefer
type
unlike

Analysis


100
75
50
25
0
%
No hypo of hypo
No new words
No pegasus
Ablation
Words
new
old

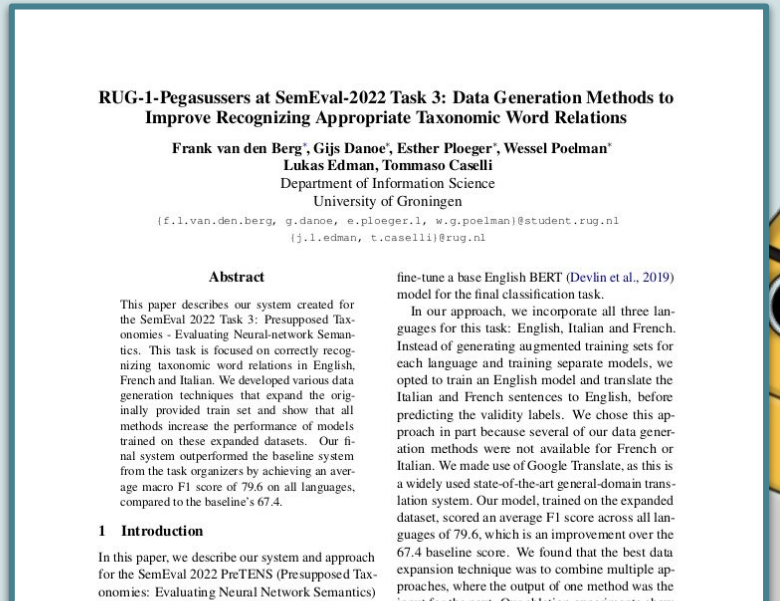

20
15
10
5
0
%
No hypo of hypo
No new words
No pegasus
Ablation
Category
animals
clothing
drinks
entertainment
food
furniture
materials
misc
people
places
plants
transportation
weather

Discussion & Conclusion

- All methods enable better performance
- Especially adding new templates, hyponyms of hyponyms, new words and paraphrasing are effective
- Using inversion is not that effective

Thank you!

PRETENS



RUG-1-Pegasussers at SemEval-2022 Task 3: Data Generation Methods to Improve Recognizing Appropriate Taxonomic Word Relations

Frank van den Berg*, Gijs Danoe*, Esther Ploeger*, Wessel Poelman*
Lukas Edman, Tommaso Caselli
Department of Information Science
University of Groningen
{f.l.van.den.berg, g.danoe, e.ploeger.1, w.g.poelman}@student.rug.nl
{j.l.edman, t.caselli}@rug.nl

Abstract

This paper describes our system created for the SemEval 2022 Task 3: Presupposed Taxonomies - Evaluating Neural-network Semantics. This task is focused on correctly recognizing taxonomic word relations in English, French and Italian. We developed various data generation techniques that expand the originally provided train set and show that all methods increase the performance of models trained on these expanded datasets. Our final system outperformed the baseline system from the task organizers by achieving an average macro F1 score of 79.6 on all languages, compared to the baseline's 67.4.

1 Introduction

In this paper, we describe our system and approach for the SemEval 2022 PreTENS (Presupposed Taxonomies: Evaluating Neural Network Semantics) fine-tune a base English BERT (Devlin et al., 2019) model for the final classification task.

In our approach, we incorporate all three languages for this task: English, Italian and French. Instead of generating augmented training sets for each language and training separate models, we opted to train an English model and translate the Italian and French sentences to English, before predicting the validity labels. We chose this approach in part because several of our data generation methods were not available for French or Italian. We made use of Google Translate, as this is a widely used state-of-the-art general-domain translation system. Our model, trained on the expanded dataset, scored an average F1 score across all languages of 79.6, which is an improvement over the 67.4 baseline score. We found that the best data expansion technique was to combine multiple approaches, where the output of one method was the

Extra slides

1. I do not like X, I prefer Y	[no hypernym relations possible] *I do not like animals, I prefer pigs.* ***[INVALID]*** *I do not like pigs, I prefer animals.* ***[INVALID]***
2. I like X, except Y	[X is superset of Y] *I like animals, except pigs.* ***[VALID]*** *I like pigs, except animals.* ***[INVALID]***
3. I like X more than Y	[no hypernym relations possible] *I like animals more than pigs.* ***[INVALID]*** *I like pigs more than animals.* ***[INVALID]*** *I like jewelry more than jazz.* ***[VALID]***
4. I like X, and more specifically Y	[X is superset of Y] *I like animals, and more specifically pigs.* ***[VALID]*** *I like pigs, and more specifically animals.* ***[INVALID]***
5. I like X, an interesting type of Y	[Y is superset of X] *I like animals, an interesting type of pig.* ***[INVALID]*** *I like pigs, an interesting type of animal.* ***[VALID]***
6. I like X, and Y too	[no hypernym rel] *I like animals, and pigs too.* ***[INVALID]*** *I like pigs, and animals too.* ***[INVALID]***
7. I like X, but not Y	[Y cannot be superset of X] *I like animals, but not pigs.* ***[VALID]*** *I like pigs, but not animals.* ***[INVALID]***

We find that there are three main template relations possible:

1. [X is superset of Y]
2. [no hypernym rel]
3. [Y cannot be superset of X]

TRAIN	TEST
Templates: I do not like X, I prefer Y. I like X, except Y. I like X more than Y. I like X, an interesting type of Y. **Nouns:** All except materials and persons	**Templates:** I like X, and more specifically Y. I like X, and Y too. I like X, but not Y. All templates with people and materials. **Nouns:** Materials Persons