AnlamVer: Semantic Model Evaluation Dataset for Turkish - Word Similarity and Relatedness

The 27th International Conference on Computational Linguistics (COLING 2018)

Gökhan Ercan, Olcay Taner Yıldız

DEPARTMENT OF COMPUTER ENGINEERING
IŞIK UNIVERSITY
ISTANBUL, TURKEY

August 24, 2018

Main Contributions

1. First word similarity and word relatedness dataset for Turkish. [1]
2. An open-source web-based word similarity questionnaire software. [2]
3. Novel analysis and visualization of semantic spaces, owing to getting bi-dimensional scores for each word-pair.
4. Dataset design considerations where the main objective is balancing word-pairs by multiple morphological and semantic attributes.

[1] Publicly available at http://www.gokhanercan.com/anlamver

[2] Publicly available at http://www.gokhanercan.com/wsquest

SIMILARITY - RELATEDNESS DISTINCTION

Types of Distributional Relations

Syntagmatic: Words co-occur at the same time.[3]
→ semantic relatedness

Paradigmatic: Words share neighbors, but not at the same time.
→ semantic similarity (e.g. synonym, antonymy)
→ most likely in the same POS. Substitutional.

	Paradigmatic relations			
Syntagmatic relations	He	likes	white	wine
	She	loves	red	roses
	Mary	enjoys	colorful	flowers

Table: Orthogonality of syntagmatic and paradigmatic relations. Table adapted from Sahlgren's work.

[3] Magnus Sahlgren. "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces". PhD thesis. Institutionen för lingvistik, 2006.

Similarity and Relatedness Distinction

Relatedness: Occur in similar contexts at the same time. Remind each others. *Ex: "gasoline - car"*

Similarity: Refer to same thing/person/concept/action. Share similar attributes. Substitutional. Occur in similar contexts but not in the same time. *Ex: "automobile - car"*

"rose - red" should be highly **related** →7,4

"rose - red" should not be **similar** →1,6

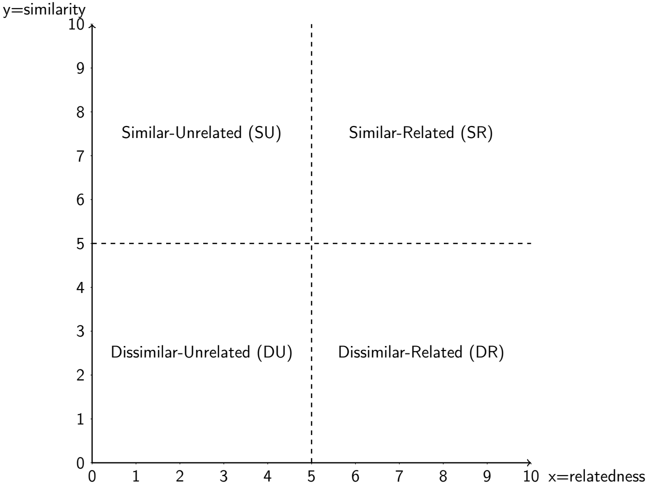
Why not having both scores at the same time?

Conventional Wordsim Datasets

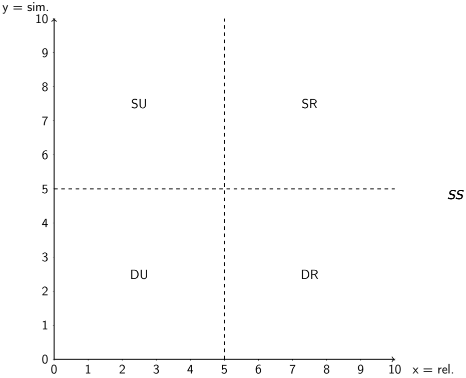
- Most *WordSim* datasets evaluates **relatedness**, not **similarity**.
- Most *WordSim* datasets lack in *clearly-defining* such distinction (WS353, RG, MC, MEN).[4] in their guidelines.
- **A "perfect" semantic model should predict two distinct scores for each word-pair.**
- Can a single model predict both?
- Decision: Getting two distinct scores for similarity and relatedness for each pair.

[4] Felix Hill, Roi Reichart, and Anna Korhonen. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation". In: *Computational Linguistics* (2016).

Sim-Rel Space: Sub-spaces

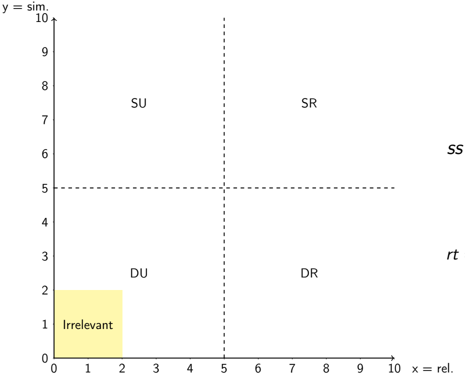


Sim-Rel Space: Sub-spaces



$$ss = f_1(r,s) = \begin{cases} \text{SU}, & \text{if } s \geq 5 \text{ and } r < 5 \\ \text{SR}, & \text{if } s \geq 5 \text{ and } r \geq 5 \\ \text{DU}, & \text{if } s < 5 \text{ and } r < 5 \\ \text{DR}, & \text{if } s < 5 \text{ and } r \geq 5 \end{cases}$$

Sim-Rel Space: Relation Types - Irrelevant



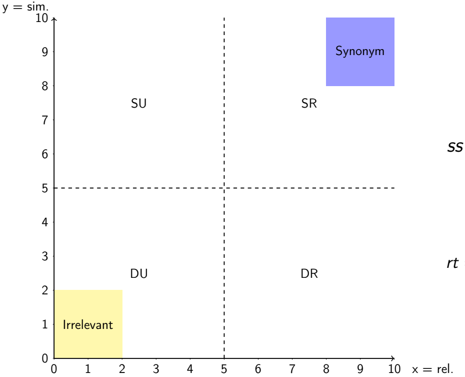
$$ss = f_1(r,s) = \begin{cases} \text{SU}, & \text{if } s \geq 5 \text{ and } r < 5 \\ \text{SR}, & \text{if } s \geq 5 \text{ and } r \geq 5 \\ \text{DU}, & \text{if } s < 5 \text{ and } r < 5 \\ \text{DR}, & \text{if } s < 5 \text{ and } r \geq 5 \end{cases}$$

$$rt = f_2(r,s) = \begin{cases} \text{irrelevant}, & \text{if } t \geq r \text{ and } t \geq s \end{cases}$$

"loose - statue"

Sim-Rel Space: Relation Types - Synonym

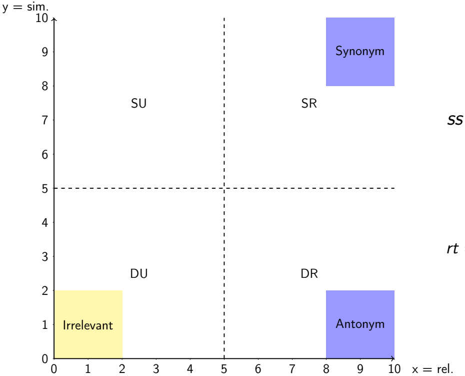
"automobile - car"



$$ss = f_1(r,s) = \begin{cases} \text{SU}, & \text{if } s \geq 5 \text{ and } r < 5 \\ \text{SR}, & \text{if } s \geq 5 \text{ and } r \geq 5 \\ \text{DU}, & \text{if } s < 5 \text{ and } r < 5 \\ \text{DR}, & \text{if } s < 5 \text{ and } r \geq 5 \end{cases}$$

$$rt = f_2(r,s) = \begin{cases} \text{irrelevant}, & \text{if } t \geq r \text{ and } t \geq s \\ \text{synonym}, & \text{if } 10 - t \leq s \text{ and } 10 - t \leq r \end{cases}$$

Sim-Rel Space: Relation Types - Antonym

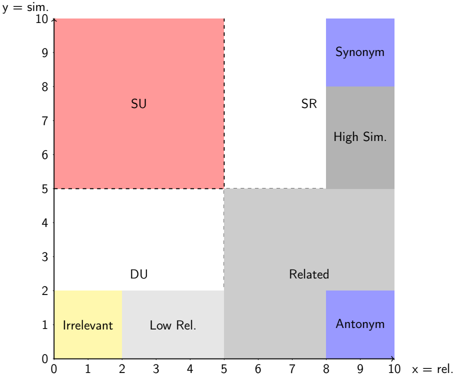


$$ss = f_1(r,s) = \begin{cases} \text{SU}, & \text{if } s \geq 5 \text{ and } r < 5 \\ \text{SR}, & \text{if } s \geq 5 \text{ and } r \geq 5 \\ \text{DU}, & \text{if } s < 5 \text{ and } r < 5 \\ \text{DR}, & \text{if } s < 5 \text{ and } r \geq 5 \end{cases}$$

$$rt = f_2(r,s) = \begin{cases} \text{irrelevant}, & \text{if } t \geq r \text{ and } t \geq s \\ \text{synonym}, & \text{if } 10 - t \leq s \text{ and } 10 - t \leq r \\ \text{antonym}, & \text{if } 10 - t \leq r \text{ and } s \leq t \end{cases}$$

"loss - profit"

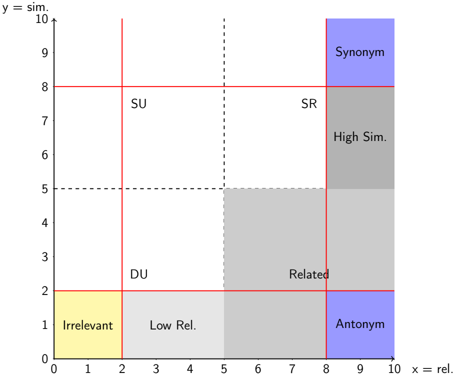
Sim-Rel Space: Similar-Unrelated (SU)



$$ss = f_1(r,s) = \begin{cases} \text{SU}, & \text{if } s \geq 5 \text{ and } r < 5 \\ \text{SR}, & \text{if } s \geq 5 \text{ and } r \geq 5 \\ \text{DU}, & \text{if } s < 5 \text{ and } r < 5 \\ \text{DR}, & \text{if } s < 5 \text{ and } r \geq 5 \end{cases}$$

$$rt = f_2(r,s) = \begin{cases} \text{synonym}, & \text{if } 10 - t \leq s \text{ and } 10 - t \leq r \\ \text{antonym}, & \text{if } 10 - t \leq r \text{ and } s \leq t \\ \text{irrelevant}, & \text{if } t \geq r \text{ and } t \geq s \end{cases}$$

Sim-Rel Space: t-Threshold



$$ss = f_1(r,s) = \begin{cases} \text{SU}, & \text{if } s \geq 5 \text{ and } r < 5 \\ \text{SR}, & \text{if } s \geq 5 \text{ and } r \geq 5 \\ \text{DU}, & \text{if } s < 5 \text{ and } r < 5 \\ \text{DR}, & \text{if } s < 5 \text{ and } r \geq 5 \end{cases}$$

$$rt = f_2(r,s) = \begin{cases} \text{synonym}, & \text{if } 10 - t \leq s \text{ and } 10 - t \leq r \\ \text{antonym}, & \text{if } 10 - t \leq r \text{ and } s \leq t \\ \text{irrelevant}, & \text{if } t \geq r \text{ and } t \geq s \end{cases}$$

TURKISH MORPHOLOGY

Turkish Morphology

- Agglutinative (Highly Inflectional and Derivational)
- 47% of word types (277K) occur only **once** in the corpus

Word	Decomposition	Sense	Frequency
maymun	maymun	monkey	very
maymunları	maymun + lAr + sH	their monkeys	medium
maymunsu	maymun + sI	ape, like monkeys	rare
maymungilleri	maymun + gil + lAr + yH	family of monkeys, primades	oov
maymuncuk	maymun + CHk	skeleton key, picklock (tool)	rare

Table: Morphological decomposition of various words sharing the same lexeme.

Problems to Address:

- OOV (out-of-vocabulary)
- RareWords

Made-up Words

Ex: "üşengeç - üşengen*" (lazy - lazy). Users scored sim: 8,2, rel: 7,8.

- Concept borrowed from phrase level model of Vecchi et al.[5].
- Even if people hear a word for the first time and it might sound odd to them, people have the intuition to make sense of the intended meaning.
- We assume that Turkish affixes can change the meanings of the words in a consistent manner, which is called *acceptable semantic deviance*.
- Our experiment showed that people can successfully understand made-up words.
- **Generalization power:** Perfect model should be able to relate made-up words as humans. Challenge for subword level models.

[5] Eva M Vecchi et al. "Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces". In: *Cognitive science* 41.1 (2017), pp. 102–136.

METHODOLOGY

Dataset Translation Issues

1. Both words in a source-pair maps to a same single word in the target language:
   Ex: ***"football - soccer"*** → ***"futbol - futbol"***
2. A word in a source-pair maps to a phrase:
   Ex: *"asylum - madhouse"* → *"tımarhane - akıl hastanesi"*.
3. Meaning loss in translations requires human re-annotation of every word-pair anyways (cross-lingual benchmarking is not possible).
4. Targeting language specific problems (OOV, rarewords). Frequency, derivations, inflections, polysemy are language dependent.

Workflow

	Stage 1	Stage 2	Stage 3
	1) Word Candidates (starts)	2) Word-Pool Selection	3) Word-Pairs Selection
Goals	1.1) Reusing existing resources	2.1) Balancing word attributes by estimations	3.1) Balancing word-pairs by estimations
Input	1.2) TKN (600) + MC (39)	2.2) Stage1 output (639) + new derivational words (99)	3.2) 320 Stage2 words
Process	1.3) Attaching frequencies, morphological tags	2.3) Filtering for balancing	3.3) Mapping pairs (every word used 2-5 times building word-pairs)
Output	1.4) 639 words	2.4) 320 words	3.4) 500 word-pairs (ends)

Stage 1: Word Candidates Selection

- Turkish word norms dataset TKN (Türkçe Kelime Normları) used. (Tekcan et al., 2005)
- Consists of 600 words annotated by 100 students.
- 480 in root form, 108 derivational, 12 inflectional.
- Has concreteness/abstractness attributes [1-7]. 'gül' is concrete (6.79), 'mutluluk' is abstract (1.85).
- Very frequent words. No OOV or rare-word based-on BOUN Corpus stats (Sak et al., 2009).

Stage 2: Word-pool Selection

- Database size target was 500 word-pairs.
- 600 words transferred from the first stage.
- Added 135 OOV and rare-words words to balance frequencies (mostly derivational).
- Grouped words in 6 frequency groups (including OOV). $(0-32, 32-320, 320-3200, 3200-32000, 32000-\infty)$.
- Frequencies numbers from Boun Corpus[6] which contains 3.2 million token types. Rare words groups defined by $gr(n, voc, g)$:

$$gr(n, voc, g) = (voc \times 10^{-(g-n+3)}) \;\&\; \texttt{"-"} \;\&\; (voc \times 10^{-(g-n+2)})$$

[6] Haşim Sak, Tunga Güngör, and Murat Saraçlar. "Resources for Turkish morphological processing". In: *Language resources and evaluation* 45.2 (2011), pp. 249–261.

Stage 2: Groupings of Word-pool

	G0	G1	G2	G3	G4	G5	Total
Frequency	OOV	RW1	RW2	RW3	RW4	RW5	
	31 9.6%	33 10.3%	30 9.3%	62 19.3%	111 34.6%	53 16.5%	320 100%
Concreteness	no value	abstract	medium	concrete			
	149 46.5%	35 10.9%	30 9.3%	106 33.1%			320 100%
Root Form	root	non-root					
	182 56.8%	138 43.1%					320 100%
Derivations	no der.	der1	der2+				
	198 61%	81 25.3%	41 12.8%				320 100%
Inflections	no inf.	inf1	inf2+				
	277 86.5%	17 5.3%	26 8.1%				320 100%

Stage 3: Word-pairs Selection

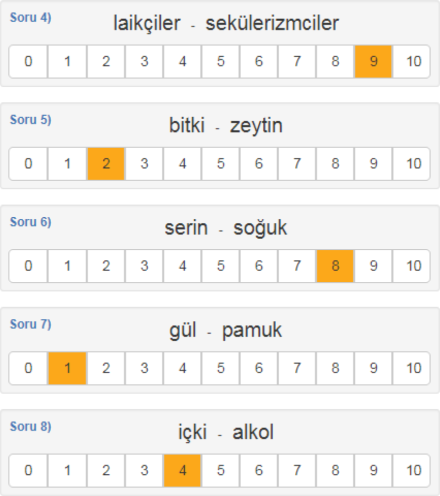
- Target: Balancing word-pair relation type ratios.
- Targeting 50 synonym, 50 antonym, 50 meronym, 50 hypernym relations.
- Pairing word manually based on our own relation type estimations.
  Ex: Paired "otomobil" and "araba" as a strong synonym candidate.
- End up with 500 word-pairs.

Methodology: Groupings of Word-pairs

	G0	G1	G2	G3	G4	G5	Total
Est. Synonyms	synonym	antonym	other				
	50 10%	50 10%	400 80%				500 100%
Est. Relatedness	high	medium	low				
	200 40%	150 30%	150 30%				500 100%
Est. Rel. Type	hyponym	meronym	other				
	50 10%	50 10%	400 80%				500 100%
OOV	no oov	any oov	two oov				
	434 86.8%	66 13.2%	42 8.4%				500 100%
Min. Derivations	no der.	der1	der2+				
	231 46.2%	166 33.2%	103 20.6%				500 100%
Min. Inflections	no inf	inf1	inf2+				
	424 84.8%	32 6.4%	44 8.8%				500 100%
Min. RareWord	rw0 (oov)	rw1	rw2	rw3	rw4	rw5	
	66 13.2%	65 13%	62 12.4%	130 26%	142 28.4%	35 7%	500 100%

QUESTIONNAIRE

Methodology: Questionnaire Design - Annotation Page



Soru 4) laikçiler - sekülerizmciler

0 1 2 3 4 5 6 7 8 [✓ 9] 10

Soru 5) bitki - zeytin

0 1 [✓ 2] 3 4 5 6 7 8 9 10

Soru 6) serin - soğuk

0 1 2 3 4 5 6 7 [✓ 8] 9 10

Soru 7) gül - pamuk

0 [✓ 1] 2 3 4 5 6 7 8 9 10

Soru 8) içki - alkol

0 1 2 3 [✓ 4] 5 6 7 8 9 10

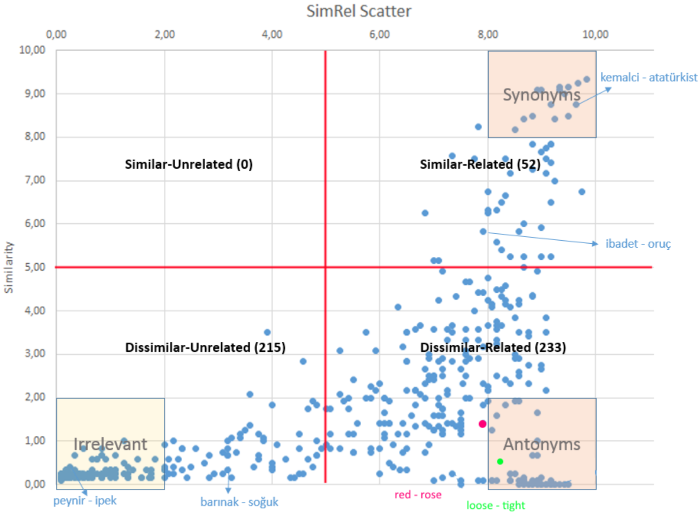
DATASET ANALYSIS

Dataset Analysis

w1	w2	avg sim	avg rel	oov	avg c.	type
otomobil	araba	9,1	9,4	no	6,87	HS,HR
üşengen	yedigen	0,5	0,1	two	-	LR,LS
kırmızı	gül	1,6	7,4	no	6,79	LS,HR
zarar	kazanç	0,18	8,8	no	3,25	ANT

- 4 participants' data removed after post-processing due to the low correlation with other participants.
- Average pairwise Spearman (ranking) correlation score: 0.748.
- Self-correlation of one participant: 0.928 (4 months between surveys)
- Lowest = 0.474, Highest: 0.847
- 0.1% null rate. Null rates replaced with average word-pair scores.

AnlamVer Sim-Rel Space Scatterplot



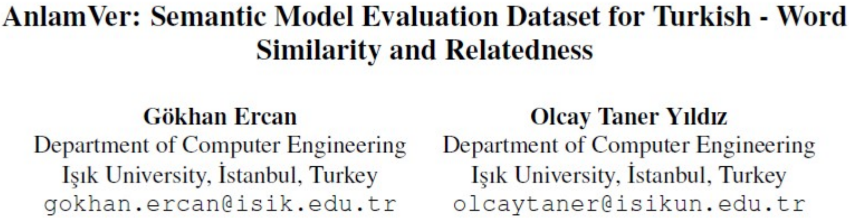
Conclusion - Possible Insights

Conventional Wordsim Dataset:
Your model's performance: %65

Proposed Dataset:

- Overall relatedness: %76, overall similarity: %36
- Abstract synonyms: %45
- Concrete antonyms: %18
- OOV performance: %32
- Irrelevants: %87
- 2+Derivations: %38
- Relatedness on SR Sub-space: %60

Thank you. Questions?



AnlamVer: Semantic Model Evaluation Dataset for Turkish - Word Similarity and Relatedness

Gökhan Ercan
Department of Computer Engineering
Işık University, İstanbul, Turkey
gokhan.ercan@isik.edu.tr

Olcay Taner Yıldız
Department of Computer Engineering
Işık University, İstanbul, Turkey
olcaytaner@isikun.edu.tr

http://www.gokhanercan.com/anlamver
http://www.gokhanercan.com/wsquest