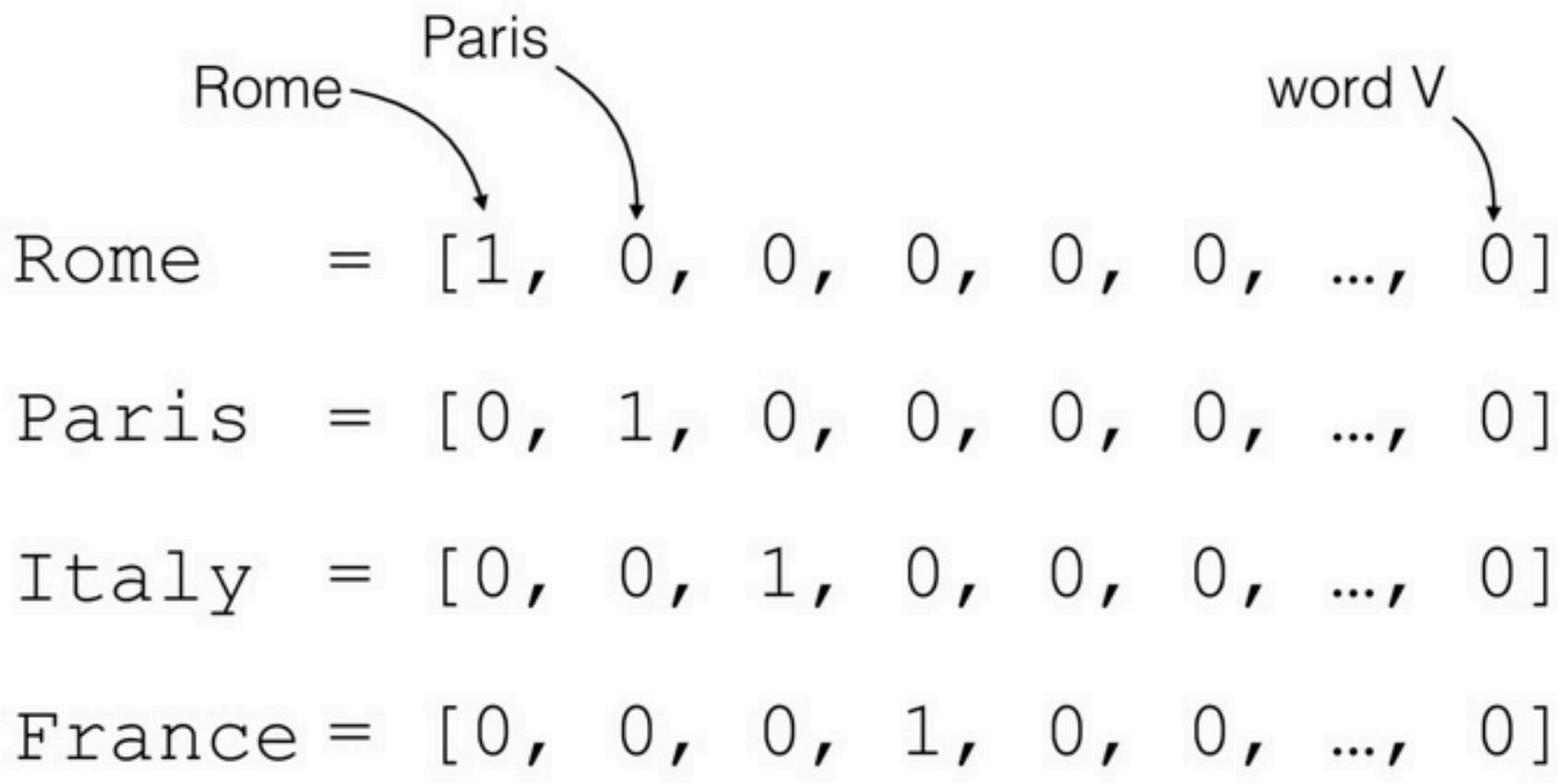
Introduction to word vectors (word embeddings)

DAAD training 2021

Representing words and meanings

- Words > Numerical representations
  - How to do this:
  - Any ideas?

One-hot encoding



Problems

- Too many ‘parameters’ to train
- No relation to word’s meaning or word relations between each other

Language in NNs: distributional similarity / "word embeddings"

- You can get a lot of value by representing a word by means of its neighbors
- "You shall know a word by the company it keeps"
  - (J. R. Firth 1957: 11)

- One of the most successful ideas of modern statistical NLP

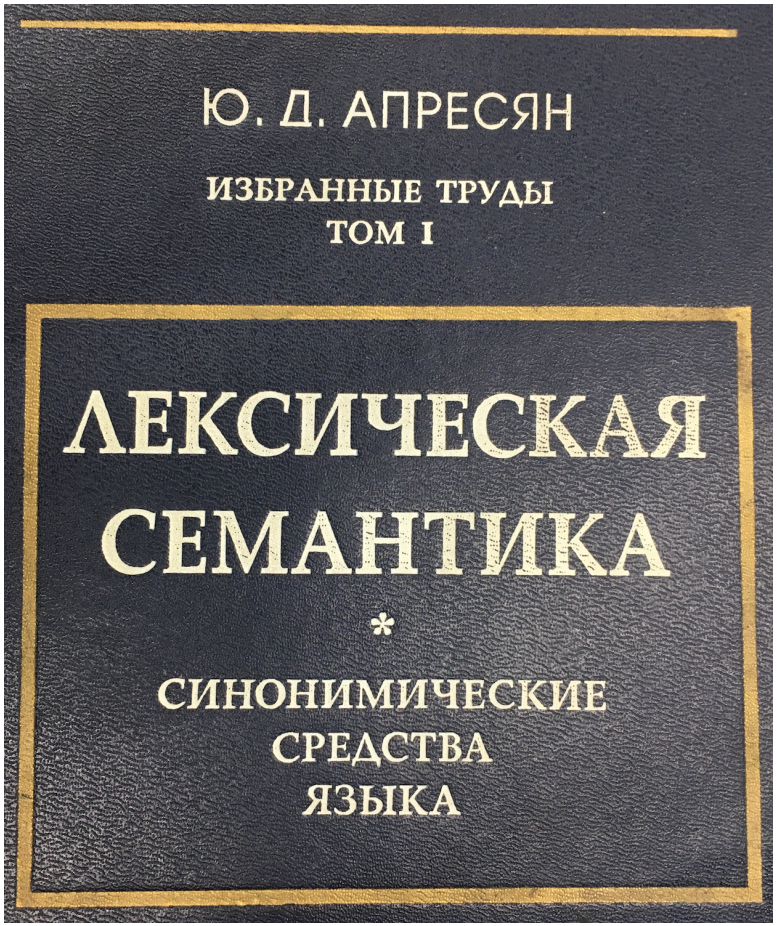
…government debt problems turning into	***banking***	*crises as happened in 2009…*
…saying that Europe needs unified	***banking***	*regulation to replace the hodgepodge…*
…India has just given its	***banking***	*system a shot in the arm…*

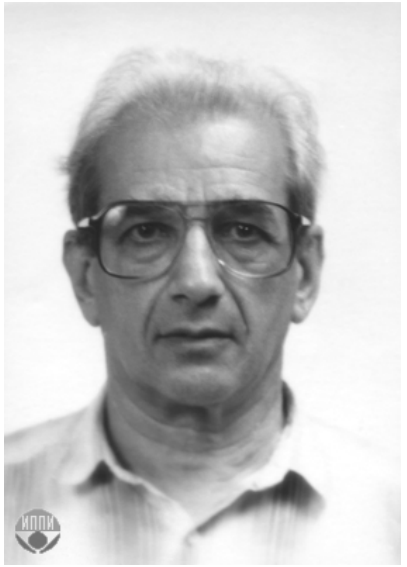
↖ These words will represent *banking* ↗

Обычное определение синонимов как слов с совпадающими или сходными значениями не опирается на строгую теорию толкований и поэтому само по себе не обеспечивает формального установления факта синонимичности–несинонимичности двух выражений. Это всегда осознавалось как серьезный недостаток теории лексических синонимов, и уже самые ранние поиски надежной операционной основы для установления факта синонимичности двух слов привели к формулировке по существу дистрибутивного критерия синонимичности – взаимозаменимости синонимов в одном и том же контексте без (заметного) различия по смыслу, хотя и с возможными стилистическими и иными различиями (см., например, Покровский 1896: 21, Блумфильд 1933: 145, Ульман 1951: 46, Курилович 1955: 74, Звегинцев 1963, Дюбуа 1964 и многие другие).

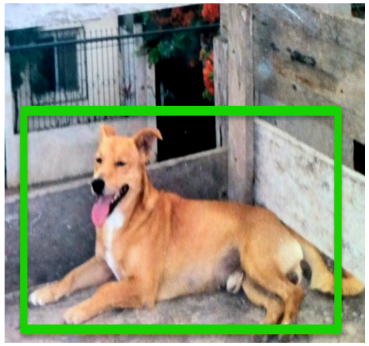
Критерий взаимозаменимости известен в двух вариантах – сильном и слабом. Сильным критерием взаимозаменимости, а именно принципом взаимозаменимости в любом контексте, оперировал в свое время С. Ульман, в некоторых своих работах определявший синонимы как слова, «идентичные по значению и взаимозаменимые в любом контексте» (цит. соч.)[5]. Очень скоро,

Более реалистичным и привлекательным казался многим исследователям слабый дистрибутивный критерий синонимичности – условие частичной взаимозаменимости синонимов в некоторых контекстах или типах контекстов (Апресян 1957, Джоунс 1964, Лайонс 1968). В этой связи заслуживают внимания идеи Дж. Лайонса. Дж. Лайонс предлагает различать а) полную—неполную синонимию (тождество – частичное сходство семантических и эмоционально-экспрессивных свойств синонимов); б) глобальную—локальную синонимию (взаимозаменимость в любых контекстах – взаимозаменимость в некоторых контекстах). В результате получается следующая классификация синонимов: 1) полные, глобальные; 2) полные, локальные; 3) неполные, глобальные; 4) неполные, локальные. Интересным свойством этой классификации является то, что в ней воплощена идея независимости совпадения—несовпадения слов по значению, с одной стороны, и их способности—неспособности к взаимозамене в одних и тех же контекстах, с другой. Правда, эта идея проведена недостаточно радикально: по крайней мере, частичная взаимозаменимость считается обязательным свойством синонимов.


Ю. Д. АПРЕСЯН
ИЗБРАННЫЕ ТРУДЫ
ТОМ I
ЛЕКСИЧЕСКАЯ СЕМАНТИКА
*
СИНОНИМИЧЕСКИЕ СРЕДСТВА ЯЗЫКА



Distributional Semantics



Dogs are man's best friend.

I saw a dog on a leash walking in the park.

His dog is his best companion.

He walks his dog in the late afternoon

…

	friend	leash	park	walking	walks	food	legs	runs	sleeps	sits	…
dog	[3	2	3	4	2	4	3	5	6	7	…]

Representing words as number sequences (semantically related words get similar representations)

word	context			
	cute	*fluffy*	*dangerous*	*of*
dog	*231*	*76*	*15*	*5,767*
cat	*191*	*21*	*3*	*2,463*
lion	*5*	*1*	*79*	*796*

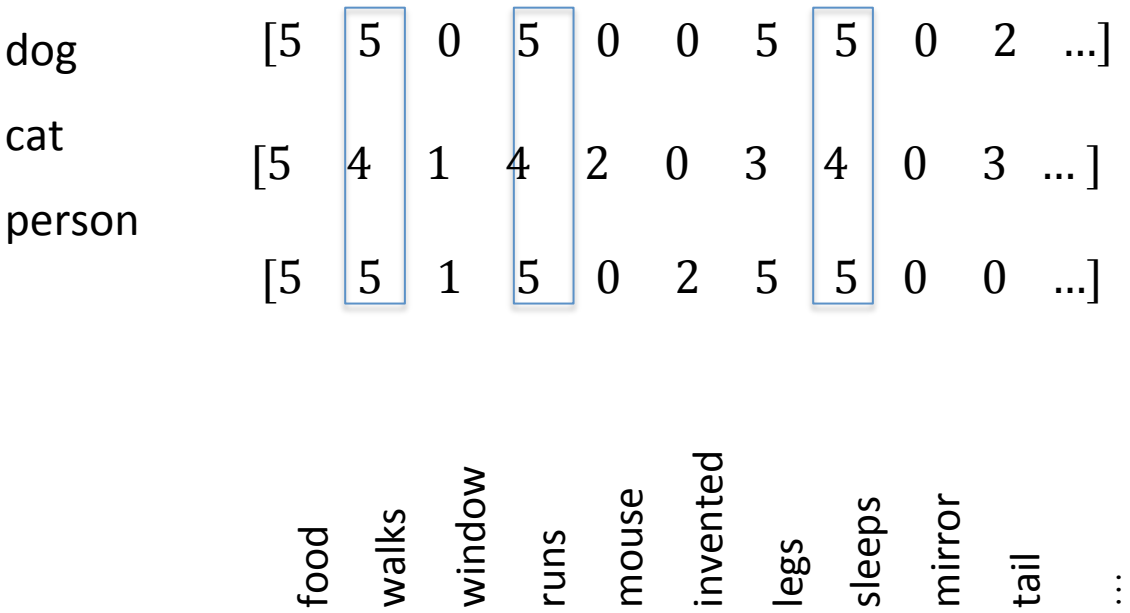
⇒

word	context			
	cute	*fluffy*	*dangerous*	*of*
dog	*9.4*	*6.3*	*0.2*	*1.1*
cat	*8.3*	*3.1*	*0.1*	*1.0*
lion	*0.1*	*0.0*	*12.1*	*1.0*

Figure 12.1 Co-occurrence statistics, such as the word *cute* occurring 231 times in the context of *dog* (left table), are converted into point-wise mutual information scores (right table). Words with similar PMI vectors (rows in the right table) have similar syntactic and semantic properties.

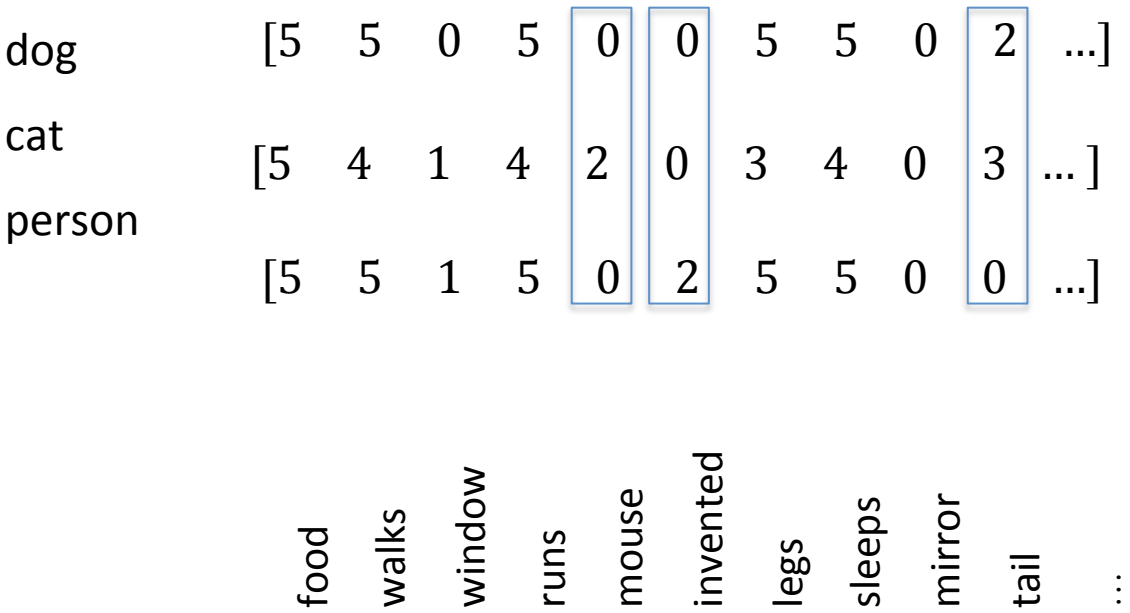
'point-wise mutual information (PMI)'

Distributional Semantics



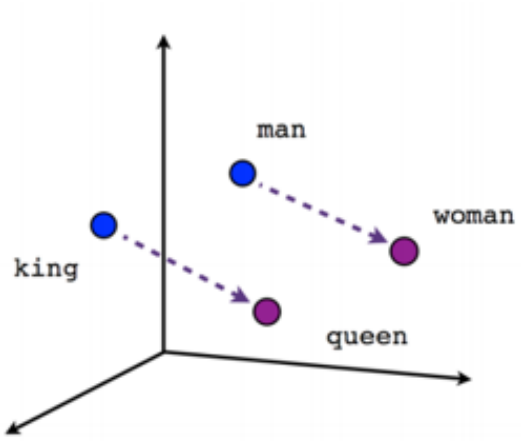
This vocabulary can be extremely large

Distributional Semantics

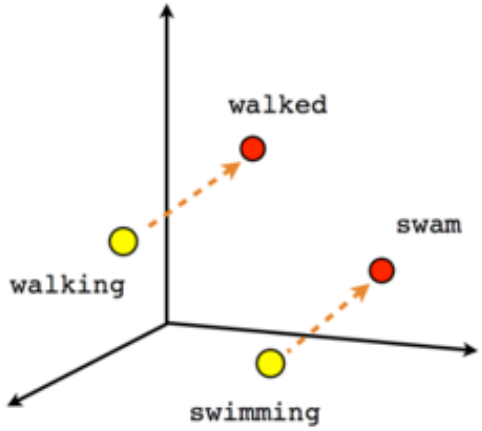


This vocabulary can be extremely large

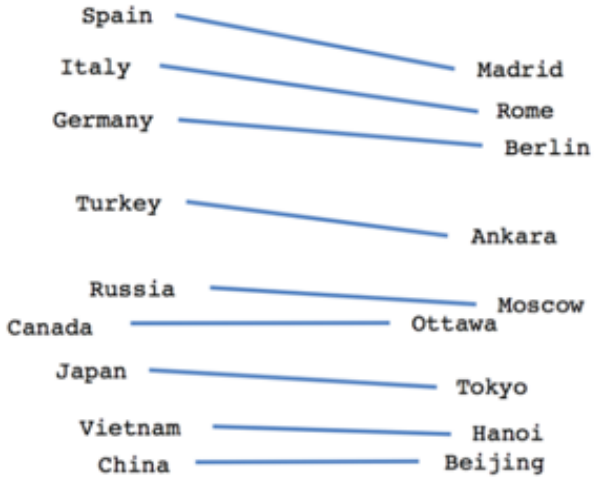
Calculation with word collocations



Male-Female



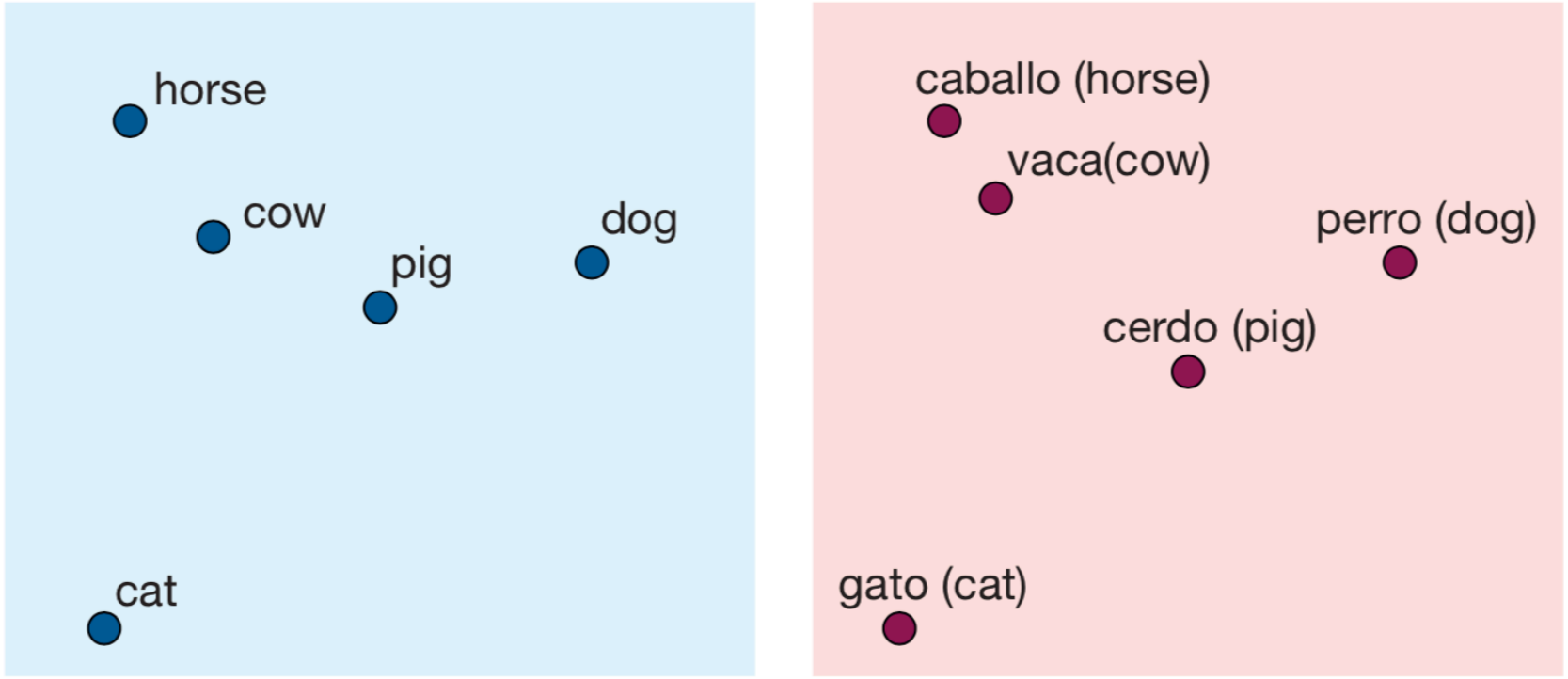
Verb tense



Country-Capital

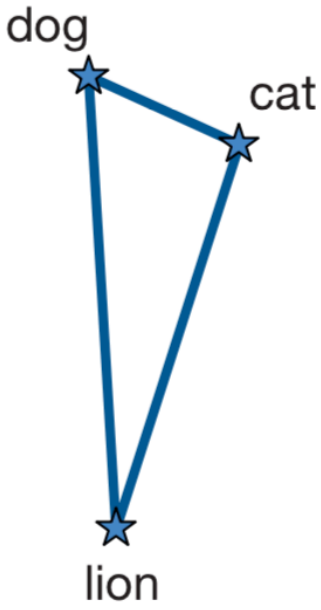
- vector[Queen] vector[King] - vector[Man] + vector[Woman]
- vector[Paris] vector[France] - vector[Italy] + vector[Rome]
  - This can be interpreted as “France is to Paris as Italy is to Rome”.

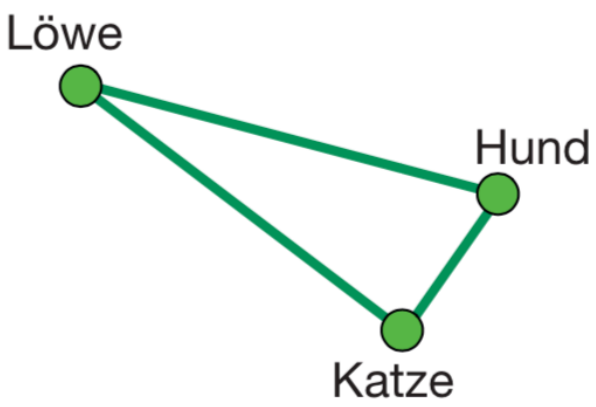
New developments: training ‘language-independent meaning representations’ on monolingual datasets

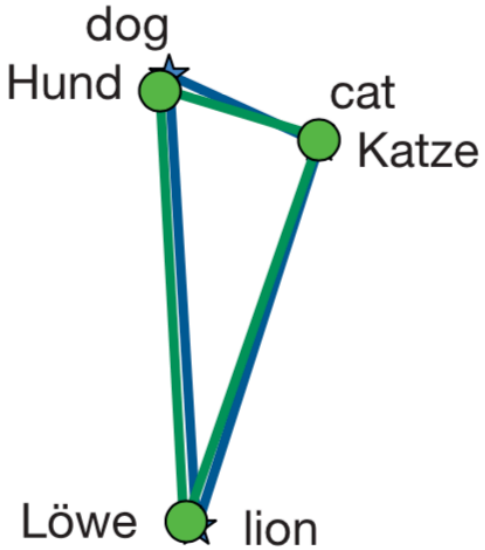


Word embedding spaces for words with the same meaning (English and Spanish),

Matching up the geometric shape of ‘word embedding’ spaces







Multiple Output Languages

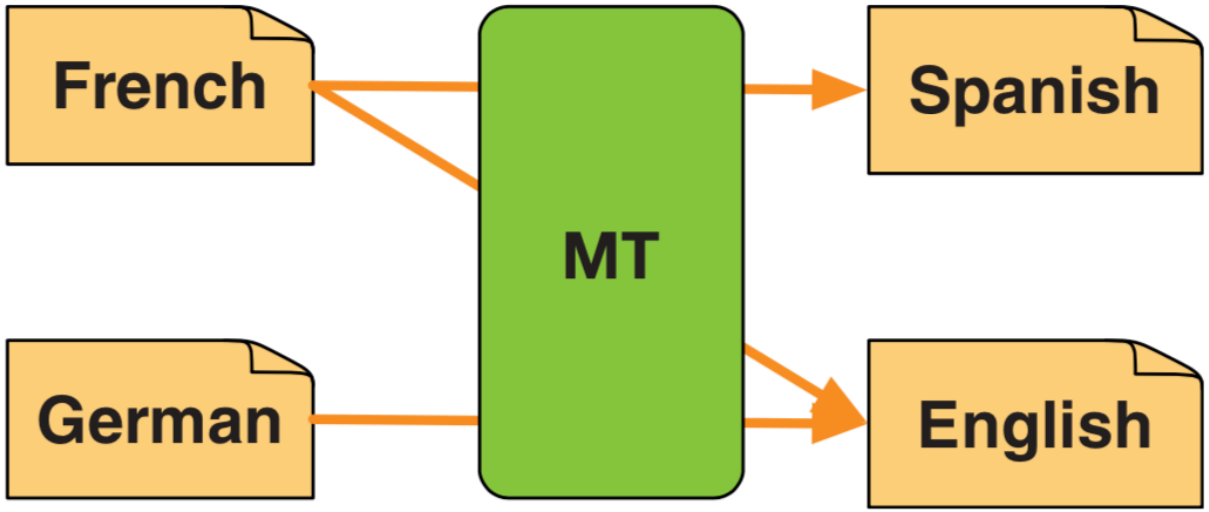
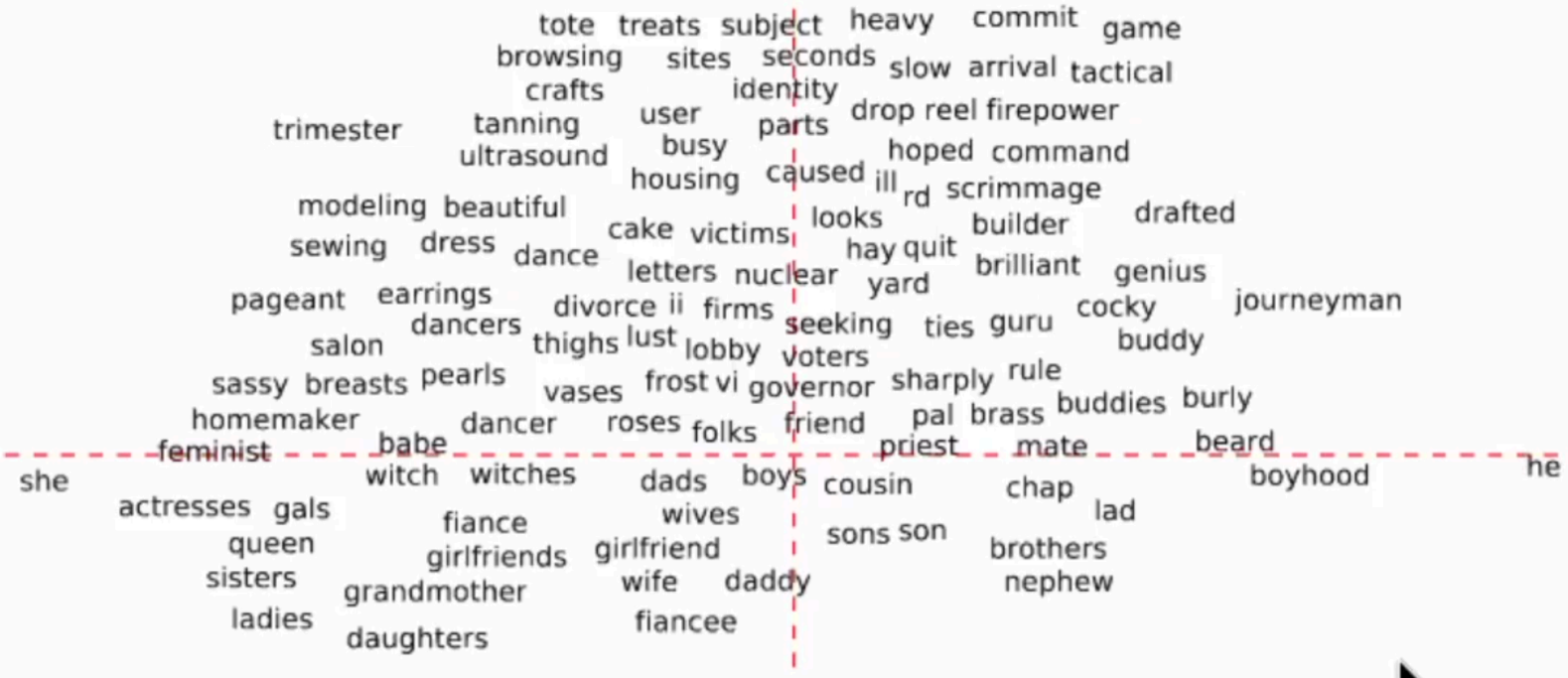


Figure 14.4 Multilanguage machine translation system trained on one language pair at a time, rotating through many of them. After training on French–English, French–Spanish, and German–English, it is possible to translate from German to Spanish.

Embeddings Encode Bias



Bolukbasi et al., "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings", NIPS 2016

Gender Bias in MT: res. By Cecilia Mahle, 2021 (IÜD Heidelberg, BA, 5th sem)

Profession – EN	**GD2 gender**		**GG2 gender**	
modifying adjective	*loving*	*ambitious*	*loving*	*ambitious*
social worker	f*	f*	f	m
office administrator	f*	f*	m	m
activist	f*	f*	m	m
professors	m*	m*	m	m
programmer	m*	m*	m	m
pilot	m*	m*	m	m
flight attendant	f*	f*	f	m
philosopher	m*	m*	m	m
receptionist	f*	f*	f	f
engineer	m*	m*	m	m
electrician	m*	m*	m	m
architect	m*	m*	m	m
dancer	f*	f*	m	m
construction worker	m*	m*	m	m
tax consultant	m*	m*	m	m
massager	f*	f*	m	m
nurse	f*	f*	f	f
doctor	m*	m*	m	m

Energy & Cost Considerations

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$T2T_{base}$	P100x8	1415.78	12	27	26	$41–$140
$T2T_{big}$	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
$BERT_{base}$	V100x64	12,041.51	79	1507	1438	$3751–$12,571
$BERT_{base}$	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

E. Strubell et al, Energy and Policy Considerations for Deep Learning in NLP, ACL 2019

Other problems with word embeddings

- Accuracy
- Interpretability
- Robustness
- Multiword expressions
- Significance for linguistic research

Other applications of word embeddings…

Application: Dynamic dictionary: decision-support system for translators

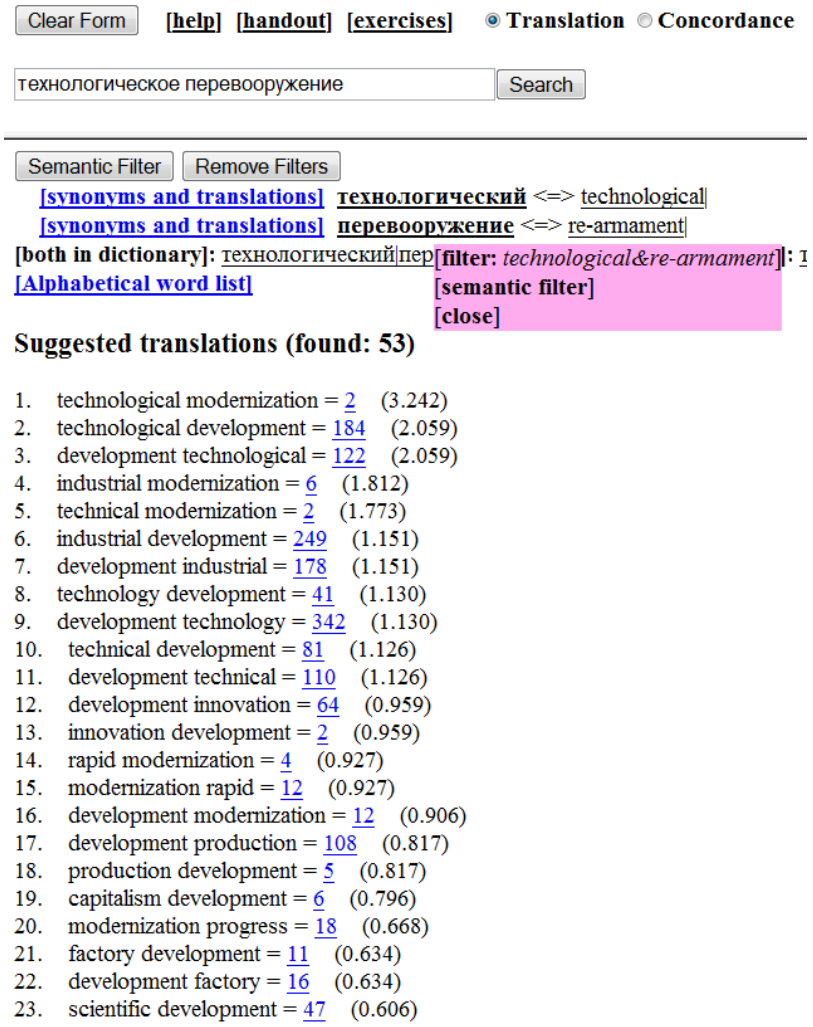
- *ASSIST* project (UK EPSRC, 2005-07)
  - Discovering indirect equivalents for MWEs in non-parallel corpora:
    - translator's amanuensis: "under the tight control of a human translator … to help increase his productivity and not to supplant him" (Kay, 1997)
  - Choice for different contexts, styles: supply multiple equivalents; authentic (not translationese)

Indirect lexical transfer

- Use of **non-compositional equivalents:**
  - *il faillit échouer* (lit.: he faltered to fail)
  - *he almost/nearly/all but failed; he was on the verge/ brink of failing/failure; failure loomed*.
  - Typically: general language expressions (not terms)
  - Open to variation
    - large number of possible translations in different contexts
  - Indirect in that they involve lexical shifts or POS transformations

Dynamic Translation Resource

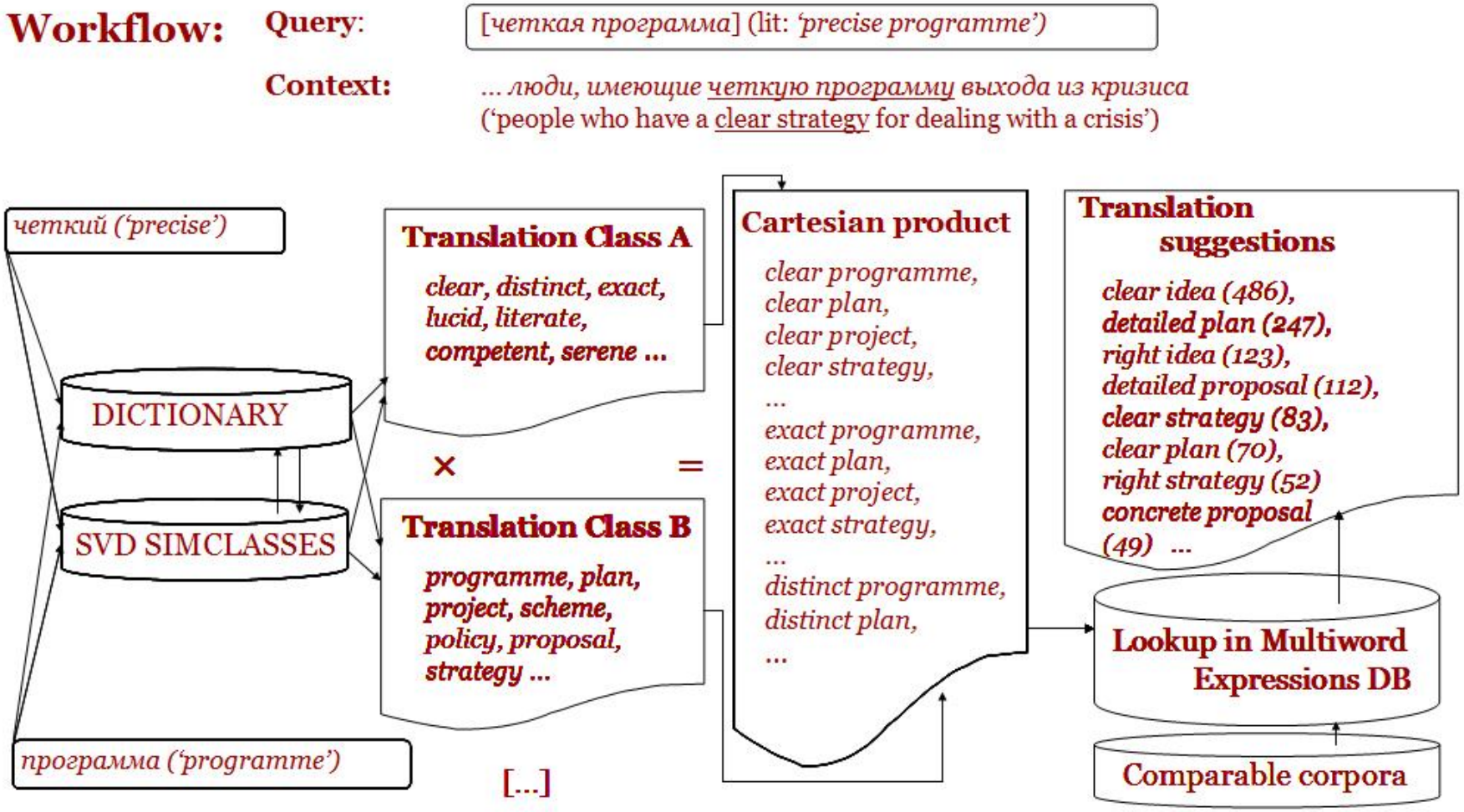
- Interface similar to a dictionary:
  - looking up single words and multiword expressions (MWEs)
- Difference: *dynamic translation resource*
  - more than an extended dictionary
  - finds equivalents for units not stored in advance
  - Searches for translation equivalents in runtime.
- Covers gaps in dictionaries and idiosyncratic MWEs (which will never be in dictionary)
- Addresses open set of translation problems



Dynamic Translation Resource: methodology

- Bilingual dictionaries (en-ru ~30k, ru-en ~50k)
- Comparable corpora: news (200M En, 70M Ru)
- Expanding the search space for each word:
  - Dictionary translations + SVD distributional similarity (Rapp, 2004)
  - *lack → absence(0.357), insufficient(0.353), inadequate(0.342), lost(0.339), shortage(0.332), failure(0.314), paucity(0.296), poor(0.288), weakness(0.282), inability(0.277), need(0.247)*
- Consistency check for multiword queries
  - computing Cartesian product between word lists
  - checking possible combinations in target corpora
    - Typically 2-4% of combinations found
- Hypothesis ranking
  - By distributional proximity to the original query
    - (Baseline: by frequency in TL corpora)
- Semantic filtering with Lancaster USAS semantic tags
  - Removes distributionally similar unrelated items

Architecture of the system



Semantic filtering: example

- Highlighted items are removed
  - *плохо отремонтированные* (badly repaired) →
    - 1. bad repair = 30 (11.005)
    - **[2. good repair = 154 (8.884)]**
    - 3. bad rebuild = 6 (5.920)
    - **[4. bad maintenance = 16 (5.301)]**
    - 5. bad restoration = 2 (5.079)
    - 6. poor repair = 60 (5.026)
    - **[7. good rebuild = 38 (4.779)]**
    - 8. bad construction = 14 (4.779)
  - *плохо (bad)* = A5– (Evaluation: Negative)
  - *good* and *well* = A5+ (Evaluation: Positive).
  - *maintenance* ontologically too far apart from *отремонтированный* (repaired)

Solutions found

- Problem: descriptors from ST as a query
  - *Дети посещают* ***плохо отремонтированные*** *школы, в которых* ***недостает*** *самого* ***необходимого***
  - (lit.: *Children attend* ***badly repaired*** *schools, in which [it] is* ***missing*** *the most* ***necessary***
- Checking transformed descriptors that fit together in TL
  - bad repair = 30 (11.005)
    bad maintenance = 16 (5.301)
    bad restoration = 2 (5.079)
    poor repair = 60 (5.026)
- Generating concordance for returned descriptors
  - (optional) useful for translation into non-native TL
  - building in **poor repair**
  - Fluent: *Children attend schools that are in* ***poor repair*** *and* ***lacking*** *basic* ***essentials***

Checking legitimate variation for translators

- Die elementarste Aufgabe des Staates
- Human translations / dictionary (1 occurrence in Europarl
  - Die ***elementarste Aufgabe*** des Staates ist der Schutz seiner Bevölkerung.
  - The most ***fundamental task*** of the state is to protect its people.
- Returned solutions:
  - essential part, crucial role, fundamental duty, essential service, basic function

Objective evaluation

- Testing on selection of indirect translation problems, extracted from a parallel corpus
  - Newspapers (R-E 118,497 wd., E-R 589,055 wd.)
- Our solutions from comparable corpora are compared with GIZA++ dictionary
  - Advantage of SMT: reuse of indirect equivalents found in parallel corpora
  - Solutions to non-compositional translation problems can be compared
- Baseline: Oxford Russian Dictionary (ORD)

Experiment set-up

- Corpus:
  - Section ~10k words of En / Ru newspapers
    - Used to train GIZA++
  - Section ~ 10k from Euronews.net – Interviews
    - Not used for training
- Tested material
  - manually selected cases of transformations
    - annotation by a human bilingual
  - converted to pairs of contextual descriptors
    - 388 from Newspapers; 174 from Interviews; 50% Ru; 50% En

Scoring Method

- Checking:
  - the ability of our system, GIZA++ dictionary and ORD to match the solution of a human translator
  - computing Recall for descriptors
- E.g.: *недостает необходимого* ([it] is missing necessary)
  - lemmatised human solution: *lack essential*
  - ORD: *missing + necessary*
  - GIZA: *missing + necessary*~0.33, *need*~0.23, *essential*~0.02
  - Our system: *lack essential* (ranked 41 without filtering and 22 with the default semantic filter)

Dynamic translation resource: evaluation ('conservative recall')

	2w descriptors		*1w descriptors*	
	news	interv	news	interv
ORD	6.70%	4.60%	32.90%	29.30%
GIZA++	13.90%	3.40%	35.60%	29.00%
Our system	**21.90%**	**19.50%**	**55.80%**	**49.40%**

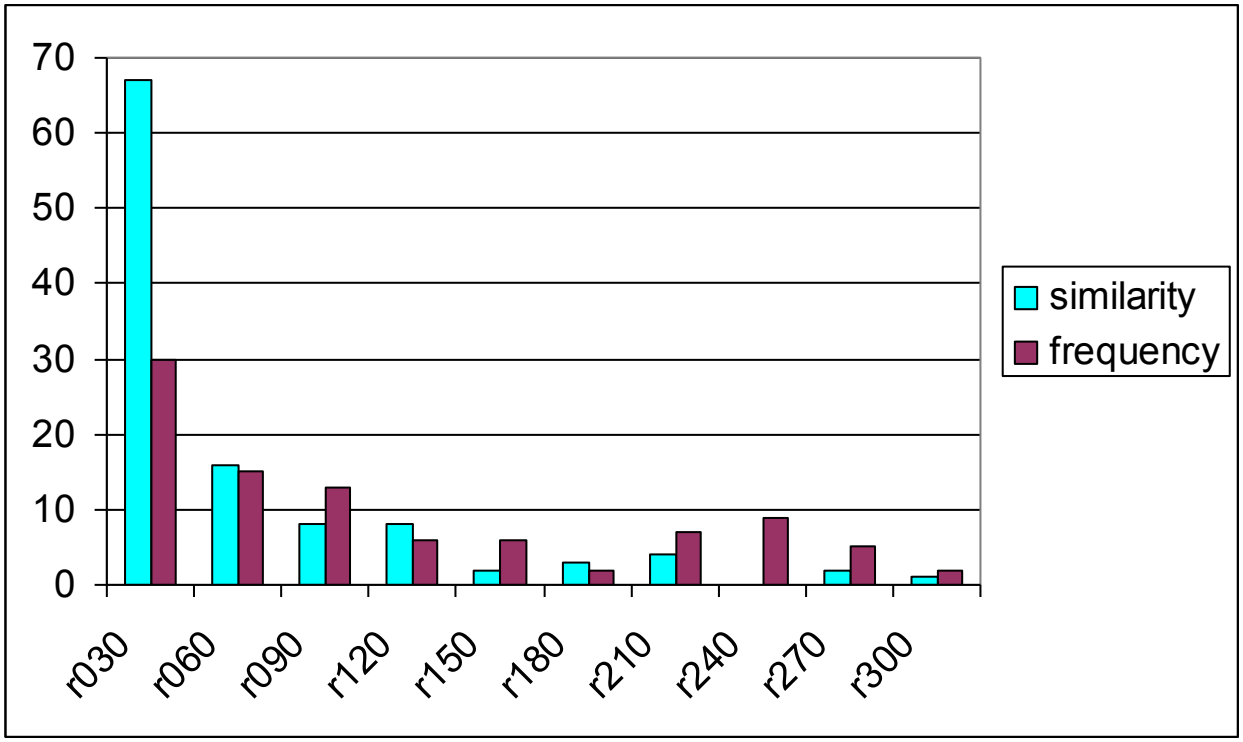
- GIZA better than ORD, drops radically on unseen text
  - equivalents in the parallel corpus are too sparse
- Our system outperforms GIZA on both corpora
  - unsupervised system: performance is stable
  - Implements 'paraphrasing' translation strategy
    - Finds solutions for previously unseen equivalents

Evaluation of hypothesis ranking

	Recall	*Average rank*
2-word descriptors		
frequency (baseline)	16.70%	rank=93.7
distributional similarity	19.50%	**rank=44.4**
sim. + semantic filter	14.40%	**rank=26.7**
1-word descriptors		
frequency (baseline)	48.20%	rank=42.7
distributional similarity	52.80%	**rank=21.6**
sim. + semantic filter	44.10%	**rank=11.3**

- Ranking by similarity: ~x2 improvement for rank
- Semantic filter: further ~x2 improvement in ranking
  - Decline in Recall, but smaller than the improvement in ranking: ~26.2% and 16.5%

Frequency polygons for rank



- Most solutions ranked by similarity appear on the first two or three screens

Subjective evaluation

- 12 professional translators
- Sentences selected from problems discussed on professional translation forums;
  - proz.com and forum.lingvo.ru
  - problems highlighted
- Tasks:
  - Find suitable suggestions produced by the tool
  - Rank usability on a scale from 1 to 5
- Experiment generated multiple reference translations

Results

- Experiment generated 210 different human solutions to 36 translation problems
- Establishing usefulness of the system for translators
  - beyond the conservative estimate
  - useful if there system output matches any of translator's solutions (≈ bleu with many refs.)
  - "Realistic" estimate of the Recall & Average Rank

	2w default	***2w with sem filt***
Conservative	32.4%; r=53.68	21.9%; r=34.67
Realistic	**75.0%; r=7.48**	**61.1%; r=3.95**

Results: human scores

	system (+)	***system (–)***
ORD (+)	4.03 (0.42)	3.62 (0.89)
ORD (–)	**4.25 (0.79)**	3.15 (1.15)

- Table: Average (and Standard Deviation) for human scores for usability
- system most useful for problems without direct dictionary equivalents, but covered by the system

Linguistic interpretation

- The tool is useful for professional translators
  - automating the search for non-trivial equivalents
- Interpretation:
  - We model the entire **family of transformations** entailed by indirect lexical transfer
    - rather than individual transformations
  - System uses a **translation strategy** based on distributional similarity in a monolingual corpus
    - applies it to novel, previously unseen examples
- Possible uses for SMT domain