Evolving New Lexical Association Measures Using Genetic Programming

Or: How to make evolution do the down and dirty work?

Jan Šnajder & Bojana Dalbelo Bašić

University of Zagreb
Faculty of Electrical Engineering and Computing
Department of Electronics, Microelectronics, Computer and Intelligent Systems

MEi:CogSci Conference
Dubrovnik, June 17, 2010

Knowledge Technologies Lab

Computer processing of language

- Natural language processing
- Computatational linguistics
- Information retrieval (IR)
- Text mining (Text analytics)

Outline

Outline

Multiword expressions

Working definition (Evert, Baldwin):

A **multiword expression (MWE)** is a combination of two or more words whose semantic, syntactic, or statistical properties are not entirely predictable from those of its components, and which therefore needs to be listed in a lexicon.

- E.g. *silver bullet*, *jump in*, *kick the bucket*, *traffic light*, *blond hair*

Three characteristic properties (Manning & Schütze):

1. non-compositionality (semantically opaque)
2. non-substitutability (lexically determined)
3. non-modifiability (syntactically rigid)

Multiword expressions (2)

MWEs cover a wide range of lexicalized expressions:

- **Idioms** (*silver bullet*, *black sheep*, *kick the bucket*)
- **Proper names** (*Humpty Dumpty*, *Star Wars*)
- **Terminological expressions** (*operating system*, *visual cortex*)
- **Phrasal verbs** (*jump in*, *call back*)
- **Compound nouns** (*traffic light*, *personal computer*, *value added tax*)
- **Institutional phrases and clichés** (*by and large*, *down and dirty*)
- ...

Important in lexicography, language learning, ...

Collocations

Our working definition:

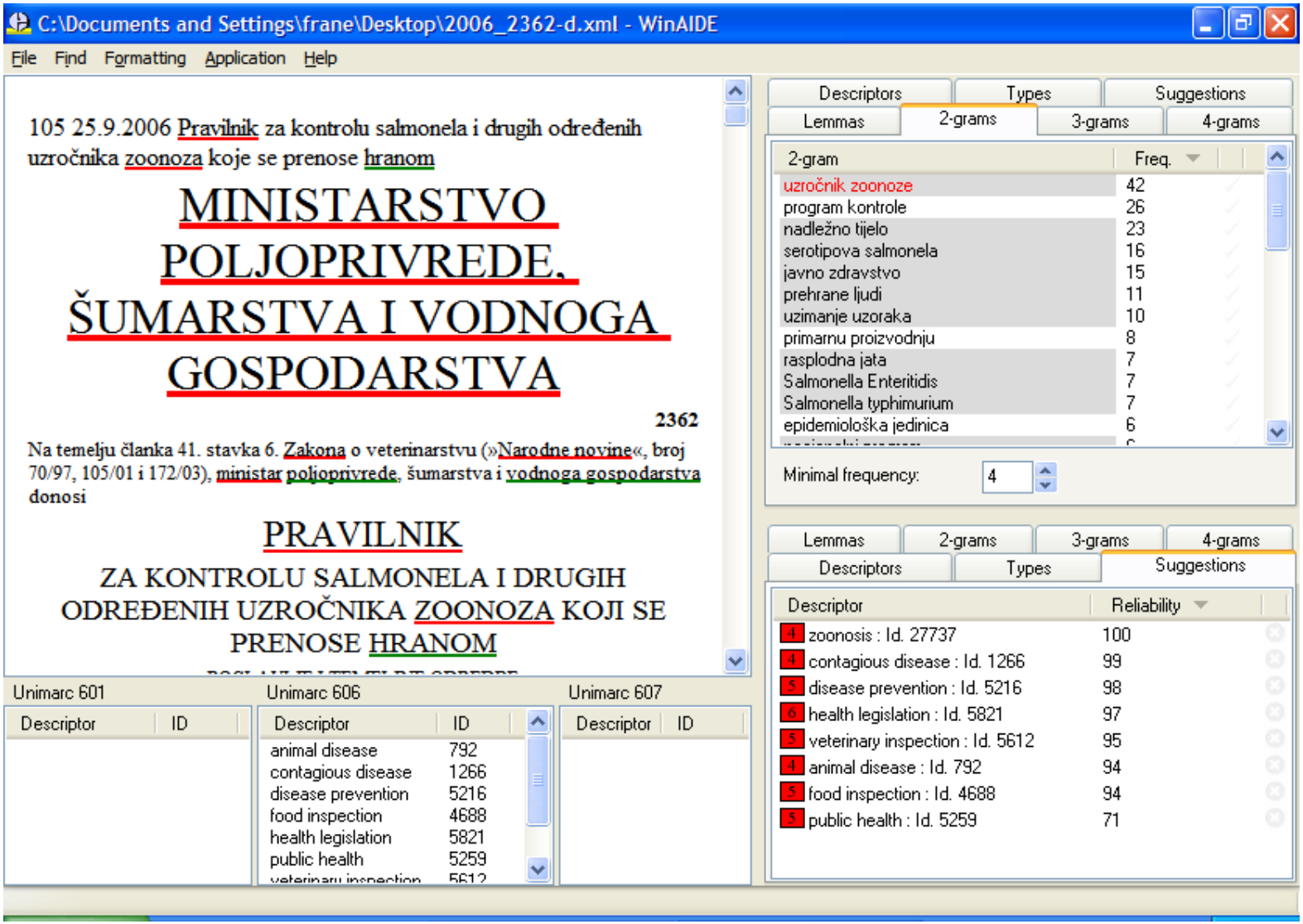
A combination of two or more words that **have the tendency to co-occur** (because they correspond to a conventional way of saying things).

- Co-occurence frequence as an epiphenomenon of lexicalization
- Collocation is an **empirical notion**, MWE is a **theoretical notion**
- Mostly overlapping
  - although some collocations are not MWE (*balloon boy*)
- Interaction between linguistics, statistics and computational linguistics

Collocations – why we need them?

- IR indexing (Vechtomova et al.; Wacholder & Songi, 2003)
- IR query expansion (Mandala et al., 2000)
- Parsing (Baldwin, 2004)
- Information extraction (Lin, 1998)
- Machine translation (Gerber & Yang, 1997)
- Natural language generation (Smadja & McKeown, 1990)
- Word sense disambiguation (Wu & Chang, 2004),
- Terminology extraction (Goldman & Wehrli, 2001)
- Document classification (Scott & Matwin, 1999)
- ...

Collocations and manual document indexing



eCADIS – Document indexing with EUROVOC descriptors (Kolar et al., 2005)

Outline

Collocation extraction from corpus

- **Idea:** use statistics to extract collocations from corpus
- **N-gram** – a sequence of two or more words (bigram, trigram, tetragram, . . .)
- The simplest approach: extract most frequent n-grams
- **Lexical association measures (AMs)** assign a value to an n-gram indicating how strong the words are associated
- AMs measure the affinity of one word towards the other: does the n-gram occur more often than expected by chance?

Example: Vjesnik (years 1999–2009) – 56MW

$f(traffic) = 12364$, $f(light) = 5741$, $f(traffic\ light) = 922$
$f(green) = 8395$, $f(dolphin) = 1067$, $f(green\ dolphin) = 2$


M·G·M
PRESENTS
GREEN DOLPHIN STREET
A Fiery Girl Who Dares The Dangers Of The Sea And A Savage Land... Fighting For The Love Of A Bold Adventurer!
M-G-M's $200,000 PRIZE NOVEL IS ON THE SCREEN!
Lana Turner
VAN HEFLIN · DONNA REED · RICHARD HART
FRANK MORGAN · DAME MAY WHITTY · EDMUND GWENN · REGINALD OWEN
GLADYS COOPER
SCREEN PLAY BY SAMSON RAPHAELSON
BASED ON THE NOVEL BY ELIZABETH GOUDGE
DIRECTED BY VICTOR SAVILLE
PRODUCED BY CAREY WILSON
A METRO-GOLDWYN-MAYER MASTERPIECE
REPRINT
R55/419

Endangered *Amazon river dolphin* (also known as the *pink dolphin*)

Lexical association measures

1. Mutual information

$$MI(x, y) = \log \frac{P(xy)}{P(x)P(y)}$$

2. Dice coefficient

$$Dice(x, y) = \frac{2f(xy)}{f(x) + f(y)}$$

3. χ^2-test

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

4. Log-likelihood

$$G^2 = 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

Extraction example

Bigram	AM
prime minister	31.5
last game	6.2
fish soup	31.0
adaptation period	1.7
pope John	22.1
new judge	13.3
least five	12.4
bread recipe	6.2
water quality	11.0
consumer basket	26.2
six euros	8.2
baloon boy	43.3
Andy Roddick	50.2
big opportunity	6.9

Extraction example

Bigram	AM
Andy Roddick	50.2
baloon boy	43.3
prime ministar	31.5
fish soup	31.0
consumer basket	26.2
pope John	22.1
new judge	13.3
least five	12.4
water quality	11.0
six euros	8.2
big opportunity	6.9
last game	6.2
bread recipe	6.2
adaptation period	1.7

Extraction procedure

1. Tokenization
2. Stemming/lemmatization (inflectional normalization)
   - *prime ministers* $\longrightarrow$ *prime minister*
3. POS tagging
   - *prime minister* $\longrightarrow$ AN
4. Token counts
5. N-gram counts
6. POS n-gram filtering
   - AN, NN, ANN, AAN, NAN, NNN, ...
7. Computing AMs for the remaining n-grams
8. Applying cutoff threshold

Heuristic AM

- Stopword-sensitive AM for trigrams (Petrović et al., 2006):

$$MI(xyz) = \begin{cases} 2 \log \frac{P(xyz)}{P(x)P(z)} & \text{if } stop(y) \\ MI(xyz) & \text{otherwise} \end{cases}$$

- Special treatment of n-grams with stopwords (propositions and conjunctions)
- E.g. *cure for cancer*, *ministry of truth*, *bed and breakfast*

Extraction evaluation

- It is important to evaluate!
- It is important to evaluate!
- It is important to evaluate!
- **Precision (P)** – how many of the extracted n-grams are genuine collocations?
- **Recall (R)** – how many of all the genuine collocations are extracted?
- Combined measure:

$$F_1 = \frac{2PR}{P + R}$$

- Need a hand-annotated **sample of collocations**
- A highly subjective task (even, or especially, for linguists)

Annotation sample

županije u iznosu
odluku Upravnog vijeća
terora pasa lutalica
Hrvatski filmski savez
veterani i kadeti
lutkarstvo Umjetničke akademije
sustav oružanih snaga
zemlje Srednjeg istoka
kriteriji za dobivanje
neposrednoj blizini mjesta
akcijskog plana vlade
beljskog pogona Poljoprivreda
drugoligaši i trećeligaši
drugih izvora financiranja
zgrada nije etažirana
vrata do vrata
sredstva za opremanje
dogradonačelnik Petar Mlinarić

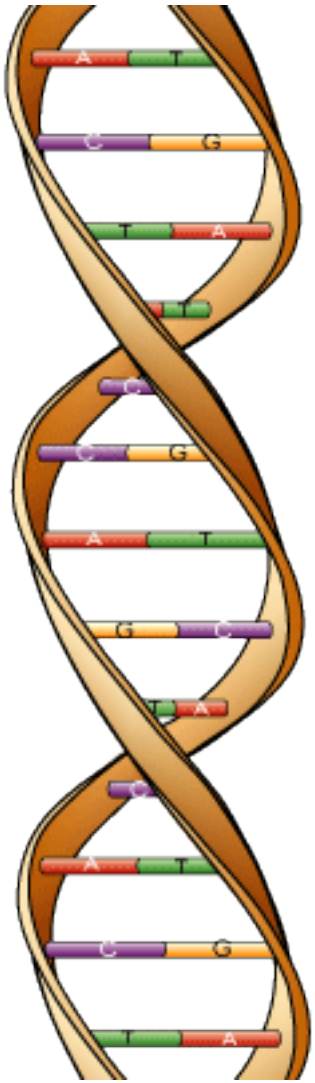
umirovljenika Ivana Matiševa
tijelima lokalne uprave
lokaciju i gradnja
mjesto u sustavu
operirano slijepo crijevo
izbora u koaliciji
groblju u Vinkovcima
činjenje takvih djela
kuna za otkup
posao za državu
klub od ispadanja
obračuna potrošnje vode
Gradsko kazalište Joza
predsjednik Uprave HT-a
stožer osječkog prvoligaša
inozemni trgovački lanci
nadogradnja i prenamjena
subote ili nedjelje

Annotation sample

županije u iznosu	umirovljenika Ivana Matiševa
odluku Upravnog vijeća	**tijelima lokalne uprave**
terora pasa lutalica	lokaciju i gradnja
Hrvatski filmski savez	mjesto u sustavu
veterani i kadeti	operirano slijepo crijevo
lutkarstvo Umjetničke akademije	izbora u koaliciji
sustav oružanih snaga	groblju u Vinkovcima
zemlje Srednjeg istoka	činjenje takvih djela
kriteriji za dobivanje	kuna za otkup
neposrednoj blizini mjesta	posao za državu
akcijskog plana vlade	klub od ispadanja
beljskog pogona Poljoprivreda	obračuna potrošnje vode
drugoligaši i trećeligaši	Gradsko kazalište Joza
drugih izvora financiranja	**predsjednik Uprave HT-a**
zgrada nije etažirana	stožer osječkog prvoligaša
vrata do vrata	**inozemni trgovački lanci**
sredstva za opremanje	nadogradnja i prenamjena
dogradonačelnik Petar Mlinarić	subote ili nedjelje

Outline

Genetic programming



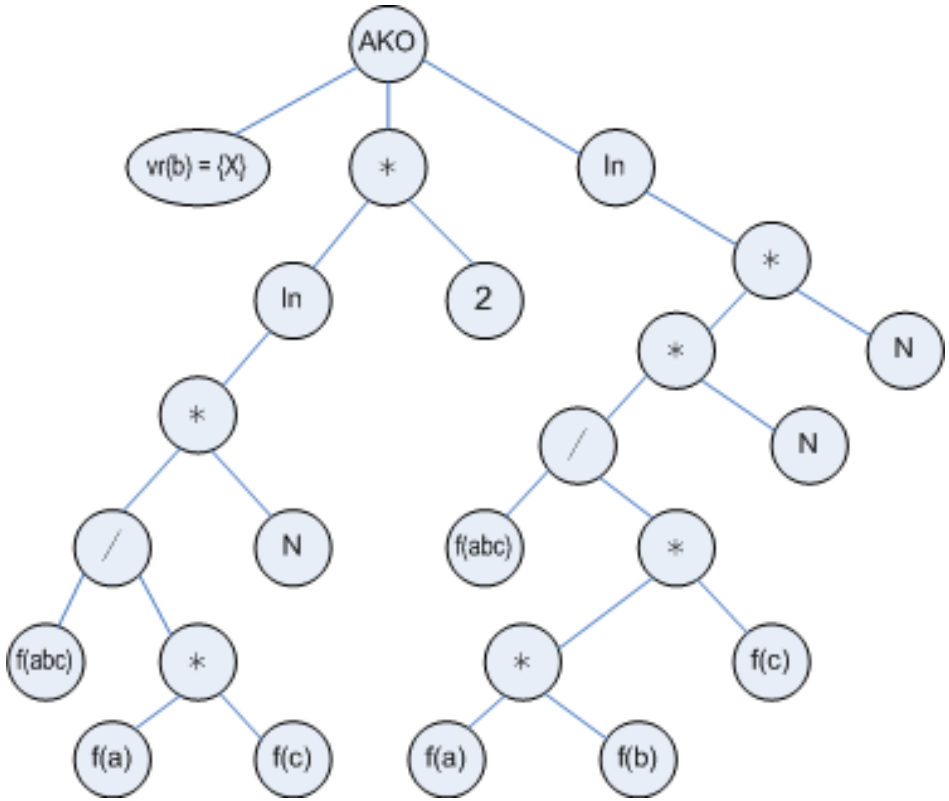
- **Genetic algorithm** – computational technique that mimics biological evolution for the purpose of solving complex optimization problems
  - population of chromosomes (solutions)
  - fitness function
  - selection
  - crossover
  - mutation
- Stochastic search through vast solution space
- **Genetic programming** – use of genetic algorithms to evolve computer programms
- Computer programms are **tree-like data structures**

Evolving AMs

Idea: use genetic programing to **evolve new AMs** (Šnajder et al., 2008)

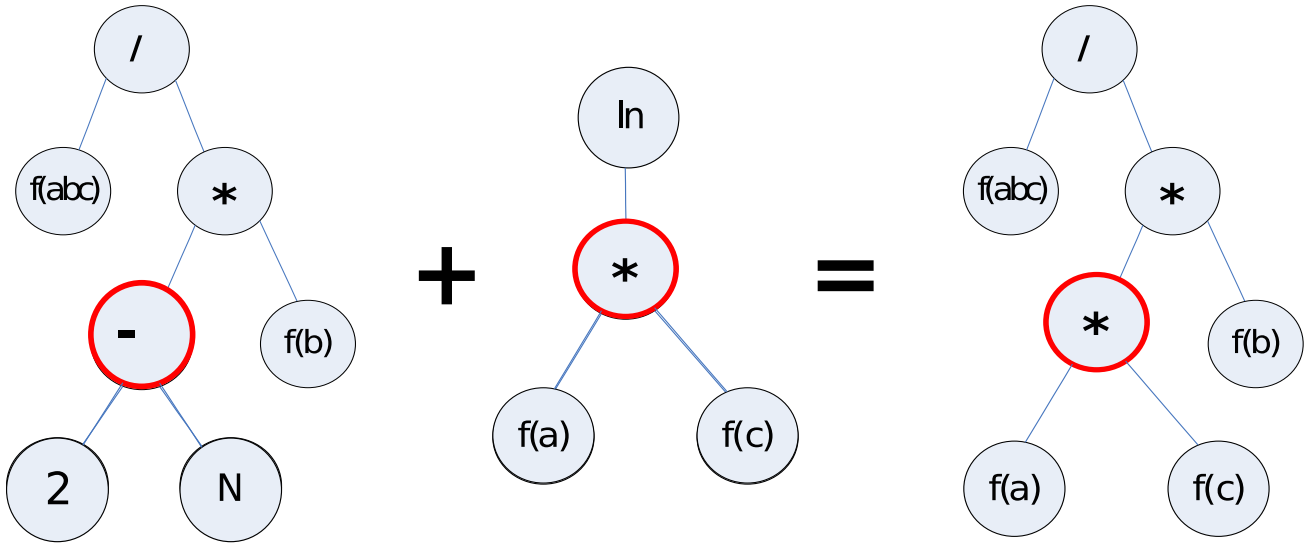
Chromosomes = AMs

- Leaves:
  - constant
  - n-gram frequency
  - POS
- Inner nodes:
  - arithmetic operator: +, -, log
  - conditional branching: "if POS is T then..."



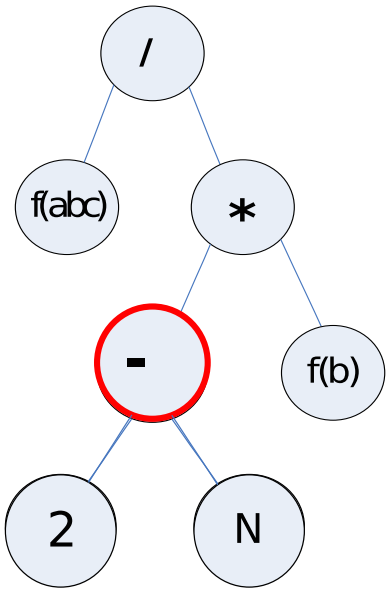
AM crossover

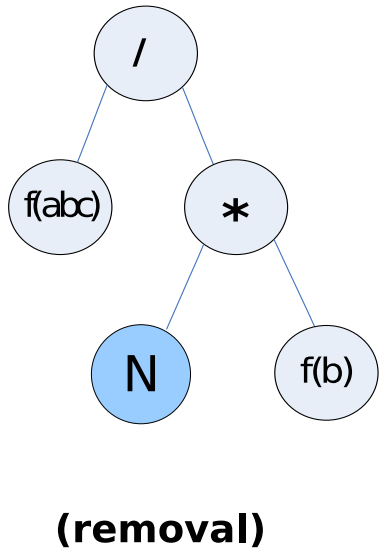
- **Exchange** of genetic material (exploitation principle)
- Two AMs (the parents) are combined into a new AM (the child)
- **Swap** randomly chosen subtrees of parent solutions

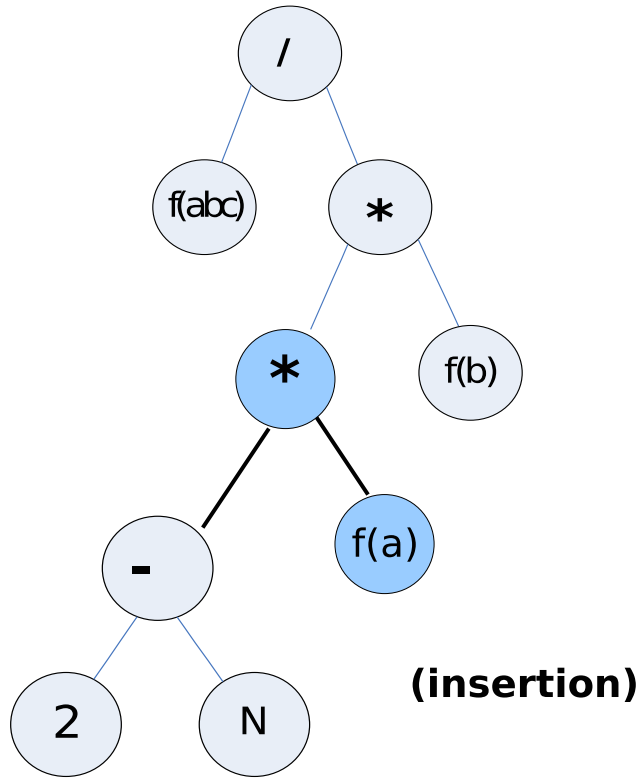


AM mutation

- **Introduce** new genetic material (exploration principle)
- **Remove** a randomly selected node (25% chance)
- **Insert** a random node at a random position (75% chance)





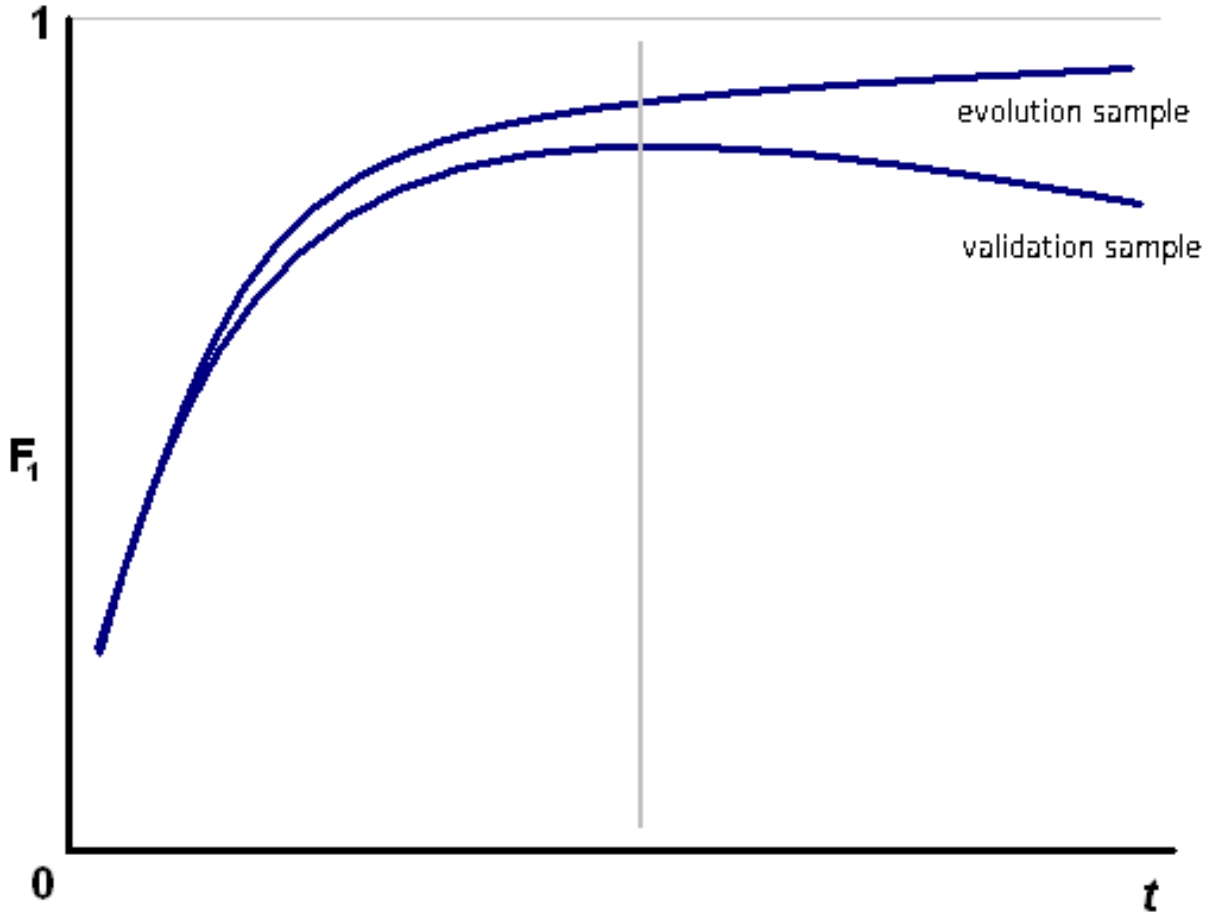


AM fitness

1. Fitness measured in terms of AM's F_1 score
   - measurd on a sample of **100 positive/100 negative** examples
2. **Parsimony pressure**
   - prevents unlimited growth of the tree
   - Ocam's razor principle (less is more)

$$fitness(j) = F_1(j) + \eta \frac{L_{max} - L(j)}{L_{max}}$$

Stopping criterion



- k iterations without fitness improvement on **validation sample**

Experiment

- **Goal:** evolve 3-gram AMs for legal corpora collocation extraction
- **Corpus:** 7008 legislative documents
- Hand-annotated samples
  - **evolution:** 100 positives/100 negatives
  - **validation:** 100 positives/100 negatives
  - **positives:** compound nouns, terminological expressions, proper names

Experiment (2)

Initial population:

- 50 – 50,000 randomly generated solutions (AM trees)
- known measures have a 50% chance of being included in the initial population

Evolution parameters:

- three-tournament selection
- parsimony η: $[0, 0.05]$
- mutation probability: $[0.0001, 0.3]$,
- 800 runs with randomly chosen parameters

Results

- **20**% measures with $\mathbf{F_1} > \mathbf{80}\%$
  - 23% if known AMs are included in the initial population, 13% if not
- Overall best: $\mathbf{F_1} = \mathbf{88.4}\%$ with 205 nodes

Measure M_{205}

```
f(abc) f(a) f(c) * / f(abc) f(ab) f(c) - f(c) f(bc) f(b)
-f(abc) + / + / N * f(b) + * ln f(c) f(b) * * N f(a) *
f(abc) f(a) f(abc) f(a) f(c) * / f(bc) * f(bc) f(b) + /
f(a) N IF(vr(b)={X}) * (-14.426000) f(b) + / N * f(bc) f(b)
-(2.000000) * ln ln / f(a) f(c) * (2.000000) * ln ln / N *
ln * / f(bc) * f(bc) f(b) + / N * (-14.426000) f(b) + / N *
f(abc) N f(a) * f(a) f(abc) f(a) f(c) * / f(bc) * f(abc)
f(b) + / N * (-14.426000) f(b) + / N * f(b) f(c) * ln ln /
f(abc) f(a) f(c) * / f(c) * ln ln (2.000000) * ln ln / N *
/ N * / N * ln f(c) * / f(a) f(b) + * ln ln f(abc) f(abc)
f(a) f(a) N IF(vr(b)={X}) (-14.426000) f(b) + * / N * / N *
ln f(c) * / f(a) f(b) + * ln ln * ln ln / f(abc) f(a) f(c)
* / f(a) f(b) + * ln ln (2.000000) * ln ln / N * ln ln
IF(vr(c)={X}) N * IF(vr(b)={X})
```

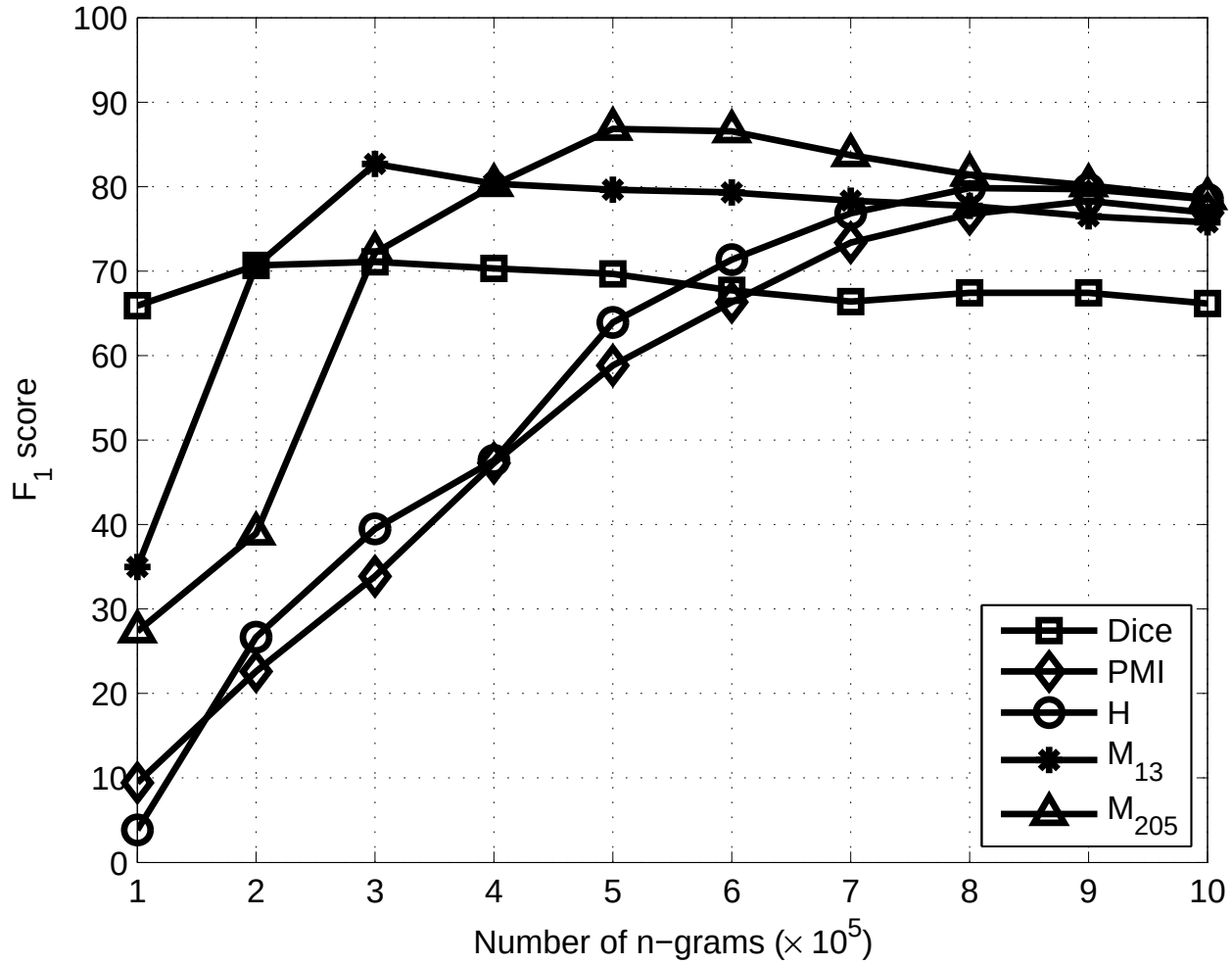
Results

- **20**% measures with $\mathbf{F_1} > \mathbf{80}\%$
  - 23% if known AMs are included in the initial population, 13% if not
- Overall best: $\mathbf{F_1} = \mathbf{88.4}\%$ with 205 nodes
- Best among solutions of less than 30 nodes:

$$M_{13}(a,b,c) = \begin{cases} -0.423\frac{f(a)f(c)}{f^2(abc)} & \text{if } stop(b) \\ 1 - \frac{f(b)}{f(abc)} & \text{otherwise} \end{cases}$$

  - evolved without measure **H** being included in the initial population!
  - 96% best-ranked measures featured a similar stopword-sensitive condition
- Confirms that trigrams with stopwords should be treated differently

Results (2)



- Measures M_{13} and M_{205} outperform the other three considered measures

Outline

Conclusion

- **MWEs** are important for NLP/IR/TM
- MWEs can be identified based on co-occurence analysis – **collocations**
- **Association measures** can be used to rank and extract collocations from corpus
- **Genetic programming** can be used to evolve new AMs
- Evolved AM will perform at least as good as any other AM included in the initial population
- Approach not limited to any specific type of collocation, language, or corpus
- FW: collocations in TM/IR, syntactic features, semantic models of collocations, ...

Selected references

- Delač, D., Krleža, Z., Dalbelo Bašić, B., Šnajder, J., Šarić, F. TermeX: A Tool for Collocation Extraction. (2009) *Lecture Notes in Computer Science (Computational Linguistics and Intelligent Text Processing).* 5449, 149–157.
- Kolar, M., Vukmirović, I., Dalbelo Bašić, B., Šnajder, J. (2005). Computer-Aided Document Indexing System. *Journal of Computing and Information Technology*, 13(4), 299–305.
- Petrović, S., Šnajder, J., Dalbelo Bašić, B. (2010). Extending lexical association measures for collocation extraction. Computer Speech and Language 24(2), 383–394.
- Petrović, S., Šnajder, J., Dalbelo Bašić, B., Kolar, M. (2006). Comparison of Collocation Extraction Measures for Document Indexing. *Journal of Computing and Information Technology*, 14(4), 321–327.
- **Šnajder, J., Dalbelo Bašić, B., Petrović, S., Sikirić, I. (2008). Evolving New Lexical Association Measures Using Genetic Programming. Proceedings of ACL-08: HLT, Short Papers, 181–184.**