MTA SZTAKI DSD

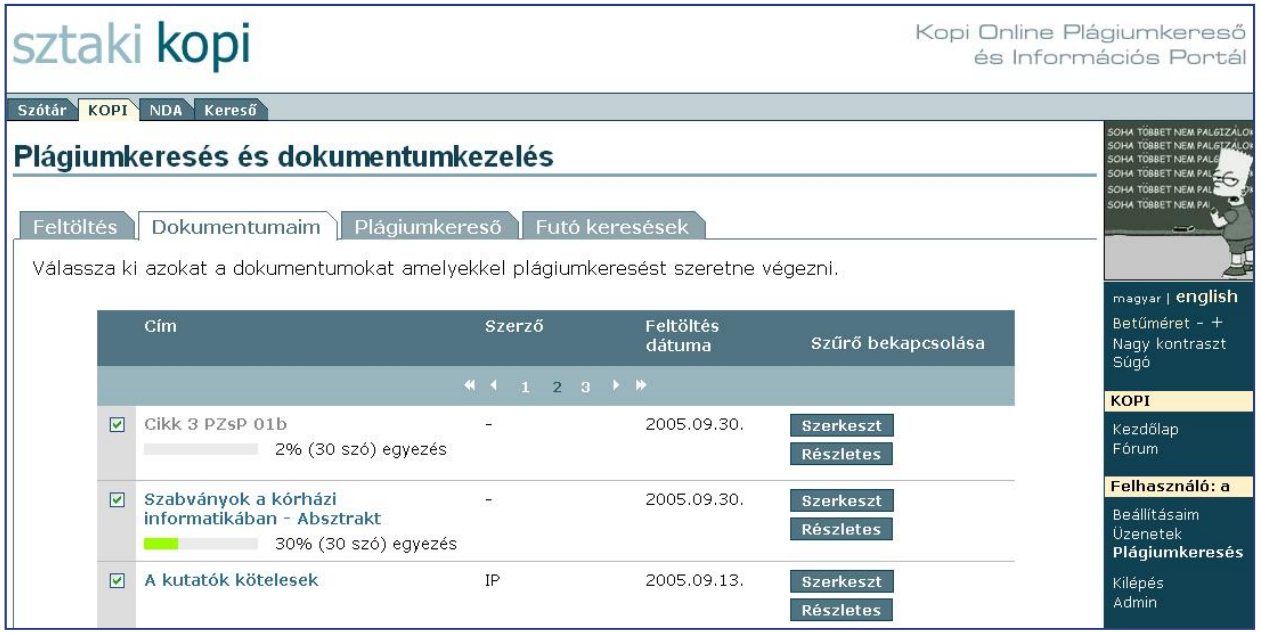
Department
of Distributed
Systems

A new approach for searching translated plagiarism

Máté PATAKI

KOPI Plagiarism Search Portal

- KOPI Online Plagiarism Search and Information Portal
- MTA SZTAKI (Computer and Automation Research Institute, Hungarian Academy of Sciences)
- http://kopi.sztaki.hu/



Problem

1. Lot of students
2. Useful information on the web
3. Theses written digitally

4. Strong foreign language skills

Problem

- Test cases for plagiarism detection software, Debora Weber-Wulff, HTW Berlin, 2010
  - 48 different plagiarism checkers
  - 42 different tests
  - *The biggest gap in all the plagiarism checkers was the **inability to locate translated plagiarism**. While this is widely expected as the technology to make such detections simply is not there.*

Problem

- CLEF 2010
- Potthast: Overview of the 2nd International Competition on Plagiarism Detection
  - *After analyzing all 17 reports, certain algorithmic patterns became apparent to which many participants followed independently. ... In order to simplify the detection of cross-language plagiarism, non-English documents in D are* ***translated to English using machine translation*** *(services).*

Other uses

- Building parallel corpora
- Searching for existing translations
- Analyzing the spread of news items
- Searching for citations
- Plagiarism detection

Why not Google translate?

- Use automatic translation engine
  - Expensive or bad quality
  - Quite poor for Hungarian
    - loose word order
    - conjugation
    - significantly different grammar
- Cross lingual document retrieval
  - Whole documents

The algorithm

- Sentence based
  - word, n-word, limb, paragraph, document

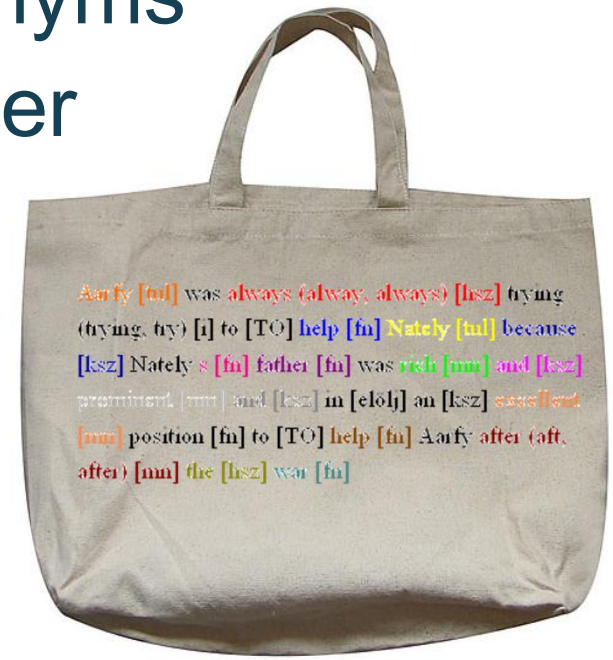
- Similarity metric

$$\mathrm{Sim}(x,y) = \min\left(\alpha \cdot |\, S_x \cap S_y \,| - \beta \cdot |\, S_x \setminus S_y \,|,\ \alpha \cdot |\, S_y \cap S_x \,| - \beta \cdot |\, S_y \setminus S_x \,| \right)$$

- "Flat" dictionary, collection of words and translations

The algorithm

- Bag of words
  - Advantages
    - no word disambiguation
    - no problem with synonyms
    - indifferent to word order
  - Disadvantages
    - large search space
    - linear search time



The algorithm

- Important parameters
  - Stopwords (by language pair)
  - Proper names and unknown words
  - Threshold (f+ / f-)
  - Size of the dictionary
  - Score for found/not found words (by word-class)

Aarfy [tul] was always (alway, always) [hsz] trying (trying, try) [i] to [TO] help [fn] Nately [tul] because [ksz] Nately s [fn] father [fn] was rich [mn] and [ksz] prominent [mn] and [ksz] in [elölj] an [ksz] excellent [mn] position [fn] to [TO] help [fn] Aarfy after (aft, after) [mn] the [hsz] war [fn]

Aarfy mindig (mindig, mind) igyekezett (igyekezik, igyekezett) Natelyn segíteni (segít) mert (mert, mer) Nately apja (apa) gazdag és befolyásos ember volt (volt, van) aki kitűnő állása (állás) révén (révén, rév) segíthetett (segít) volna (van) Aarfyn a háború után

Test environment

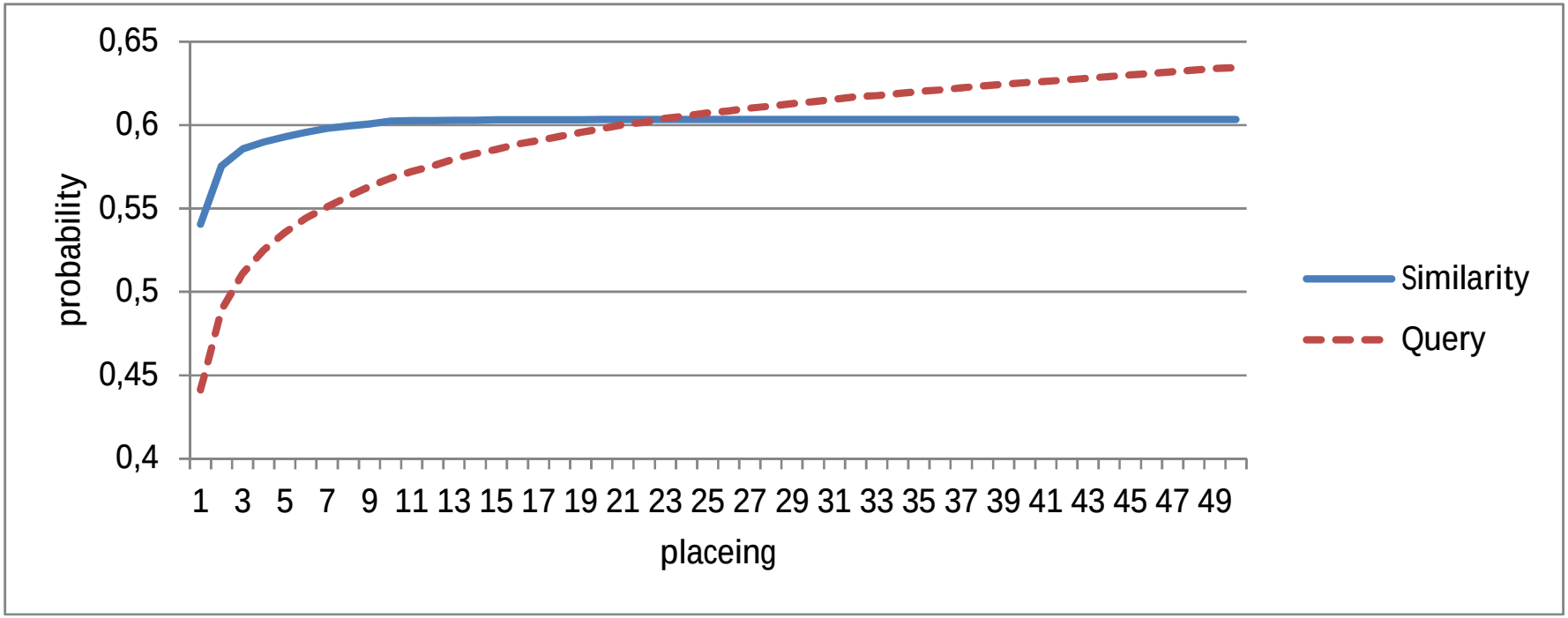
- English Wikipedia
  - 31GB XML
  - 3 800 000 articles
  - SZTAKI Desktop GRID
  - Available as a text: **http://kopiwiki.dsd.sztaki.hu/**

- Google Translate
  - For the research

Demo

- http://www.wikipedia.org
- http://translate.google.com
- http://kopi.sztaki.hu

Results



Probability of being ranked above a certain value

Results

y	*x=1*	*x=2*	*x=3*	*x=4*	*x=5*
1	0,4				
2	0,64	0,16			
3	0,784	0,352	0,064		
4	0,8704	0,5248	0,1792	0,0256	
5	0,92224	0,66304	0,31744	0,08704	0,01024
6	0,953344	0,76672	0,45568	0,1792	0,04096
7	0,972006	0,84137	0,580096	0,289792	0,096256
8	0,983204	0,893624	0,684605	0,405914	0,17367
9	0,989922	0,929456	0,768213	0,51739	0,266568
10	0,993953	0,953643	0,83271	0,617719	0,366897

Probabilities to find at least *x* out of *y* sentences

Demo

sztaki kopi

Szótár | KOPI | NDA | Kereső

Test

You can test the plagiarism search system by inserting poems of the famous Hungarian poet Sándor Petőfi. The quotes should be at least twenty words long. You can find poems of Petőfi on this page: http://mek.oszk.hu

Home

On this page: Services, Our users, Where does it search, How to use KOPI, History, Contact

Welcome to KOPI Plagiarism Search Portal!

KOPI - The best choice for searching translated plagiarism

"plagiarism: the act of using another person's words or ideas without giving credit to that person : the act of plagiarizing something" (Merriam-Webster)

Services

Monolingual plagiarism search

magyar | english
Fontsize - +
High contrast
Help

KOPI

Home
SZDG
Plagiarism search
Upload
My documents
Plagiarism search
Running searches
Messages
Forum

User: a

My settings
Logout

Demo

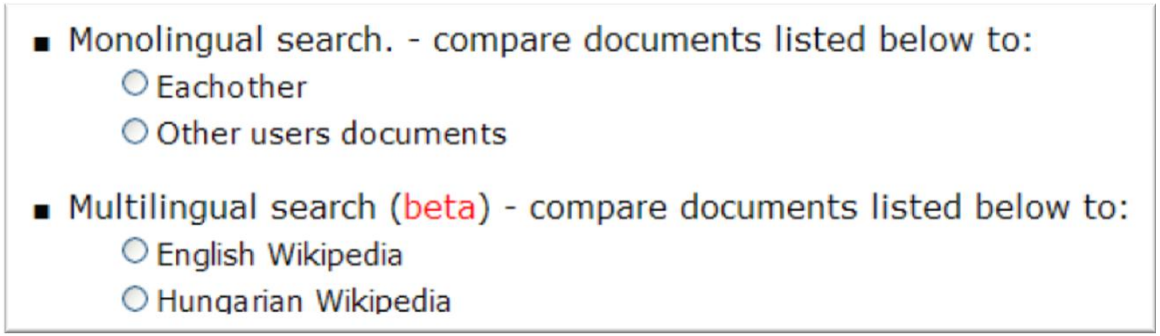
1.


Title:
Author:
Document:
Browse...
Upload

2.


KOPI Protection instead of Copy Protection
Pataki Máté
January 8, 2009
edit
detailed
Plagiarism Search Within One Document
Pataki Máté
January 8, 2009
edit
detailed

3.


Monolingual search. - compare documents listed below to:
Eachother
Other users documents
Multilingual search (beta) - compare documents listed below to:
English Wikipedia
Hungarian Wikipedia

Demo

From: KOPI
Date: 2012.01.24.
Subject: 1 dokumentum összehasonlítása az angol Wikipédiával.

[Üzenet törlése]

2 hasonló mondatot talált a rendszer 3 Wikipédia cikkben:

1. Rövidítés (3)
   Rövidítésnek (latinul abbreviatura) nevezzük közszavak és tulajdonnevek rövidített formáit, melyek szinte kizárólag írott formában élnek, azaz amelyeket kiejtve teljes alakjukban használunk.

   - rövidítés és mozaikszó egy szó, kifejezés vagy név rövidített formája
     megjegyzés 1: rövidítésnek n
     formáit, melyek szinte kizáról
     teljes alakjukban használunk.

   (utca), km (kilométer), É (észak

   - (utca), km (kilométer), É (ész

1. Pete Seeger (7)
   Seeger was born in French Hospital, Midtown Manhattan, the youngest of three sons.

   - Pete Seeger Manhattan közepén, a Midtown-nak is hívott városrész francia kórházában született.

   His father, Charles Louis Seeger Jr. was a prominent musicologist, composer, and music professor.

   - Apja, ifj. Charles Louis Seeger, zeneszerző és zenetudós volt, aki az elsők között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekből fakadó zenét.

   His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the twentieth century.

   - Nevelőanyja, Ruth Crawford Seeger egyike volt a huszadik század legkiemelkedőbb női zeneszerzőinek.

Results

HUN: A Carnatic ez? - robbant ki.
ENG: Am I on the Carnatic?" -1

HUN: A detektívnek minden oka megvolt, hogy így okoskodjon.
ENG: The detective was not far wrong in making this conjecture. -2

HUN: A detektív is hasztalanul fáradozott, hogy ő legyen a nézeteltérésben a főszereplő.
ENG: As vainly did the detective endeavor to make the quarrel his. -3

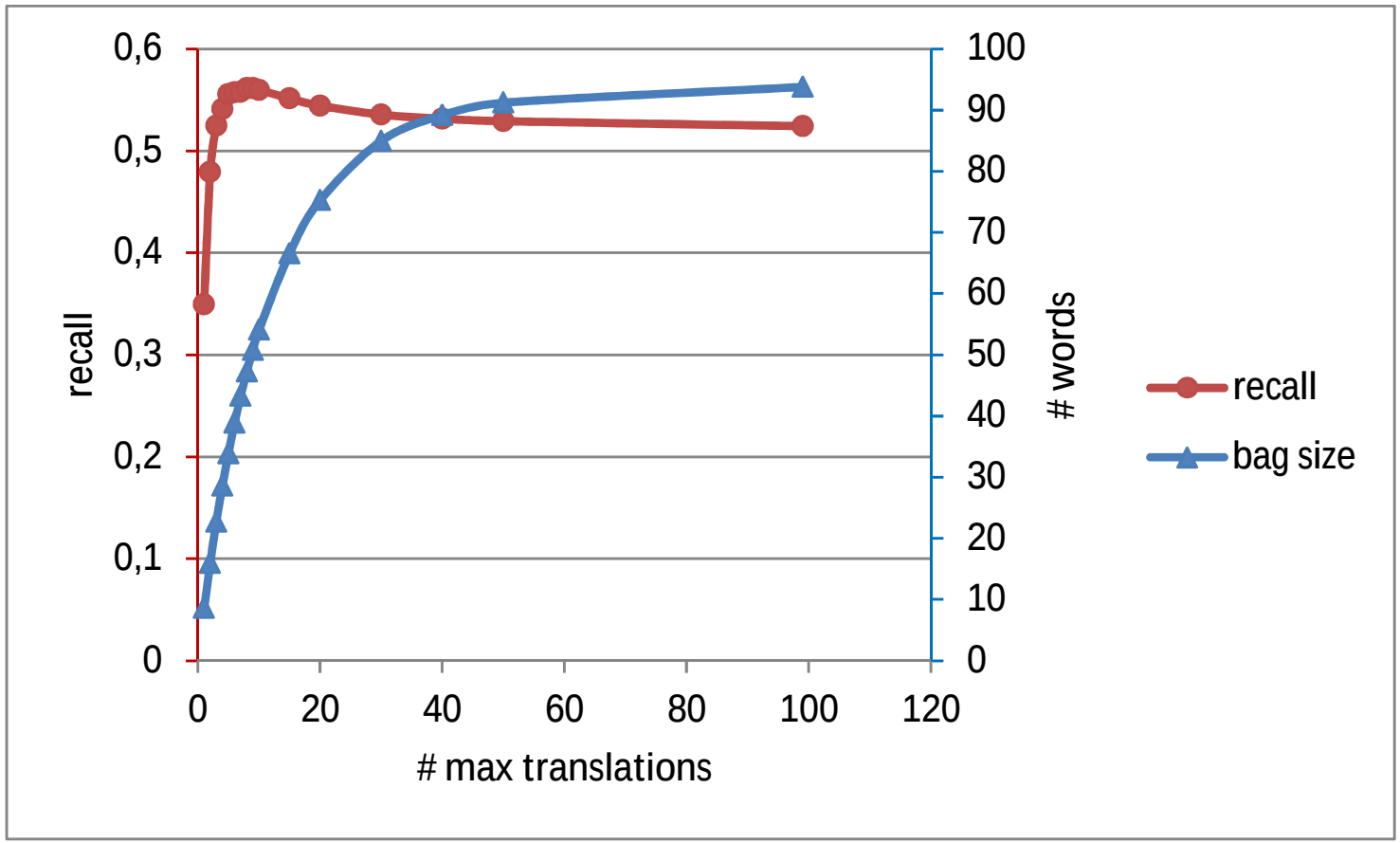
HUN: - A hídon.
ENG: "On the bridge." 2

HUN: - Addig óvadék ellenében szabadlábra helyezem mindkettőjüket.
ENG: "Meanwhile, you are liberated on bail." -3

HUN: A gentlemannek különben is kész volt a terve a továbbiakra.
ENG: Mr. Fogg's course, however, was fully decided upon. -6

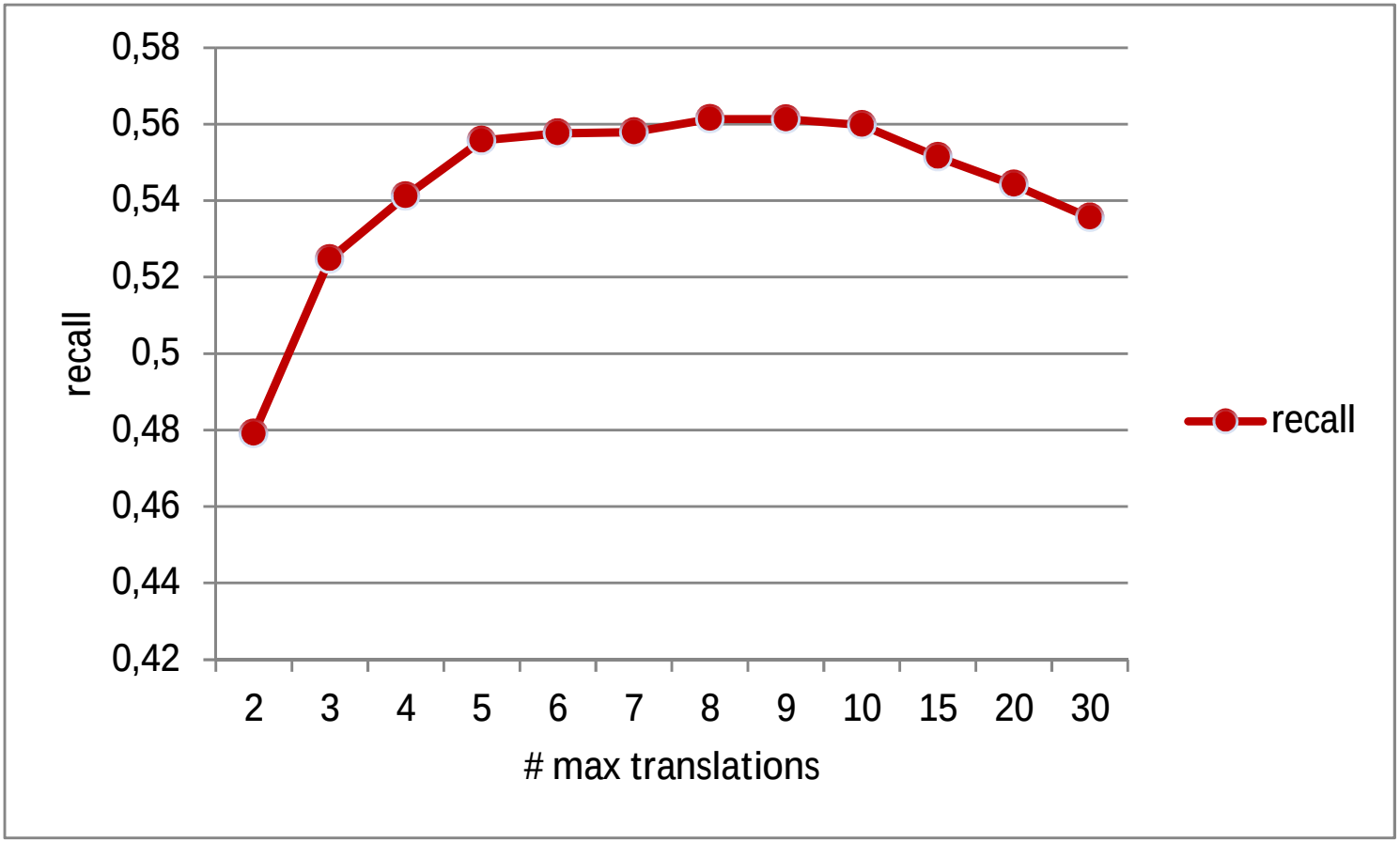
HUN: A gépész azonban történetesen épp e napon felment a fedélzetre, megkereste Mr. Foggot, és meglehetősen élénk vitát folytatott vele.
ENG: On this day the engineer came on deck, went up to Mr. Fogg, and began to speak earnestly with him. -4

Statistics



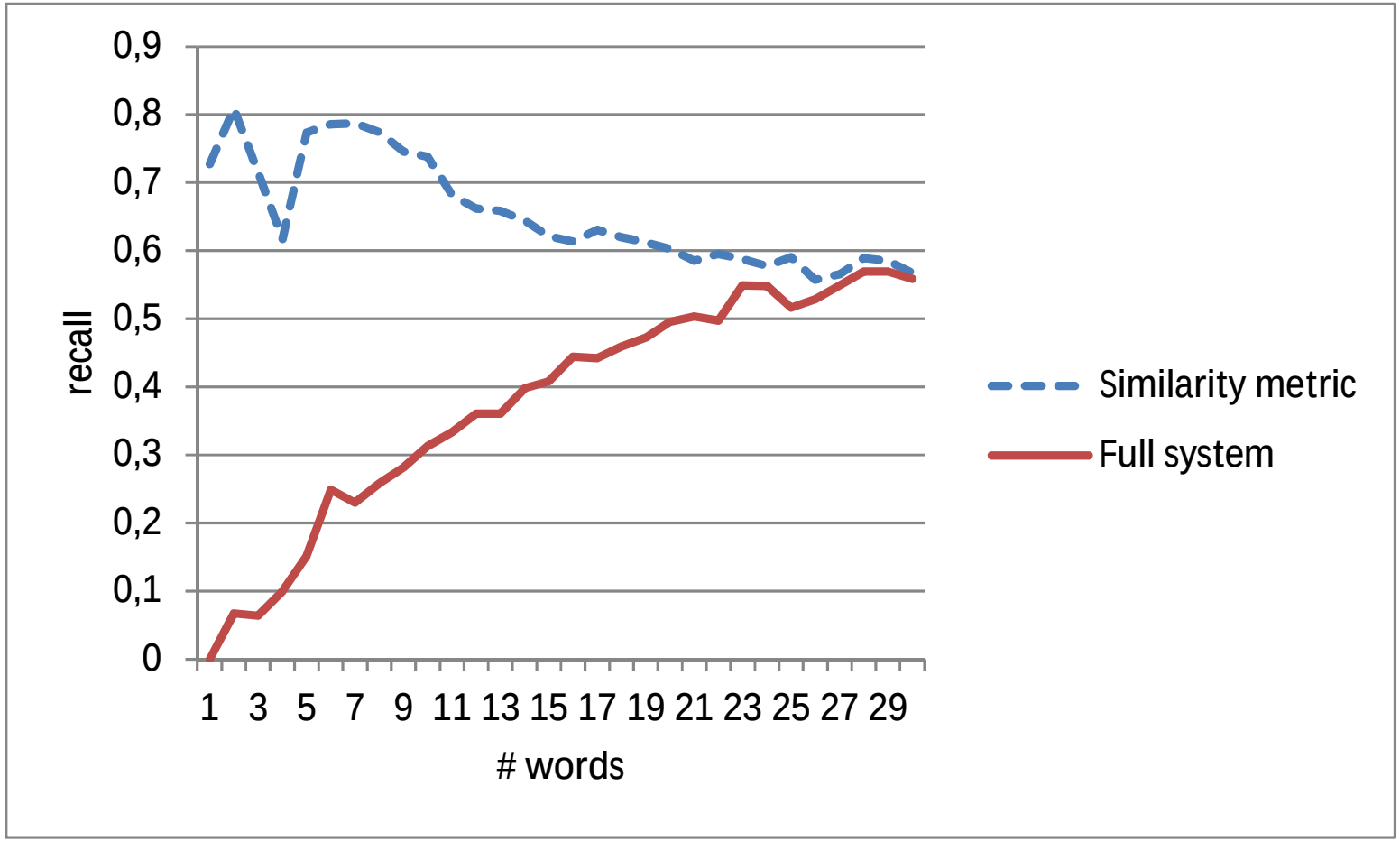
Number of words and recall as a function of the number of translations per word

Statistics



Recall as a function of the number of translations per word

Statistics



The recall as a function of sentence length

Actuality

- **End of 2011 our multilingual plagiarism detection service started**
- **2012:**

Hungary President Schmitt quits in plagiarism scandal

Hungary's President Pal Schmitt says he is resigning, after being stripped of his doctorate over plagiarism.

Mr Schmitt, elected in 2010, said "my personal issue divides my beloved nation rather than unites it".

"It is my duty to end my service and resign my mandate as president," he told parliament.

Last week, Budapest's Semmelweis University revoked his 1992 award after finding that much of his thesis had been copied.

Mr Schmitt, 69, won gold medals for fencing at the 1968 and 1972 Olympic Games.

Mr Schmitt was an Olympic fencing champion before his rise in politics

Romanian prime minister accused of plagiarism

Allegations prompt questions about government's ability to tackle misconduct in academia.

Quirin Schiermeier

18 June 2012 | Updated: 20 June 2012

Romania's new government, still reeling from a misconduct scandal that forced its research minister to resign last month, has been hit by fresh allegations of plagiarism that strike at the very top.

Prime Minister Victor Ponta has been accused of copying large sections of his 2003 PhD thesis in law from previous publications, without proper reference. If the charges are substantiated, they could spark public pressure for Ponta to resign, say political insiders. The allegations are also raising fresh doubts about the government's ability to tackle corruption in the higher-education system.


Romanian Prime Minister Victor Ponta has been accused of copying large amounts of his PhD thesis from other sources without proper attribution.

XINHUA/PHOTOSHOT

- **Added Hungarian-French within two days**

KOPI Portal

http://kopi.sztaki.hu

Recap

- Translated plagiarism detection
- Alternative method to using machine translation
  - information retrieval + a new cross-language similarity metric
- Quick to add new language pairs
- Works for monolingual search as well

Web: *http://dsd.sztaki.hu*

Email: *Mate.Pataki@sztaki.hu*