Technologies for Reusing Text from the Web

The Oral Exam of

Martin Potthast

To Obtain the Academic Degree of

Dr. rer. nat.

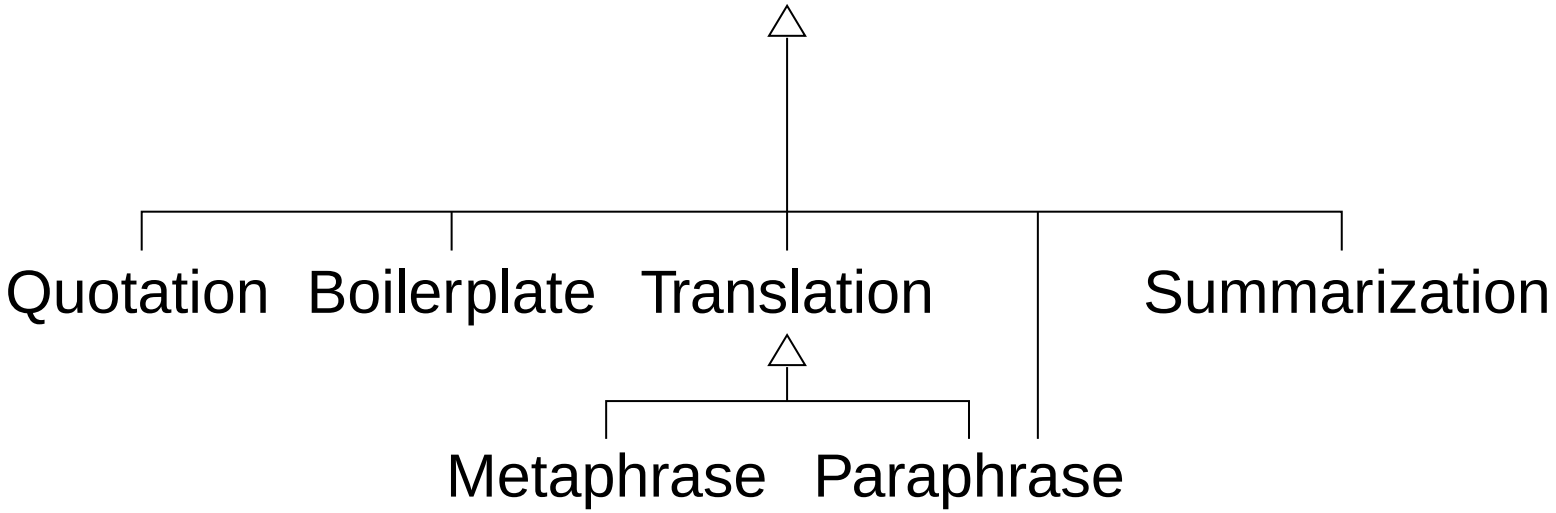
Web Technology & Information Systems Group
Bauhaus-Universität Weimar

www.uni-weimar.de www.webis.de www.potthast.net

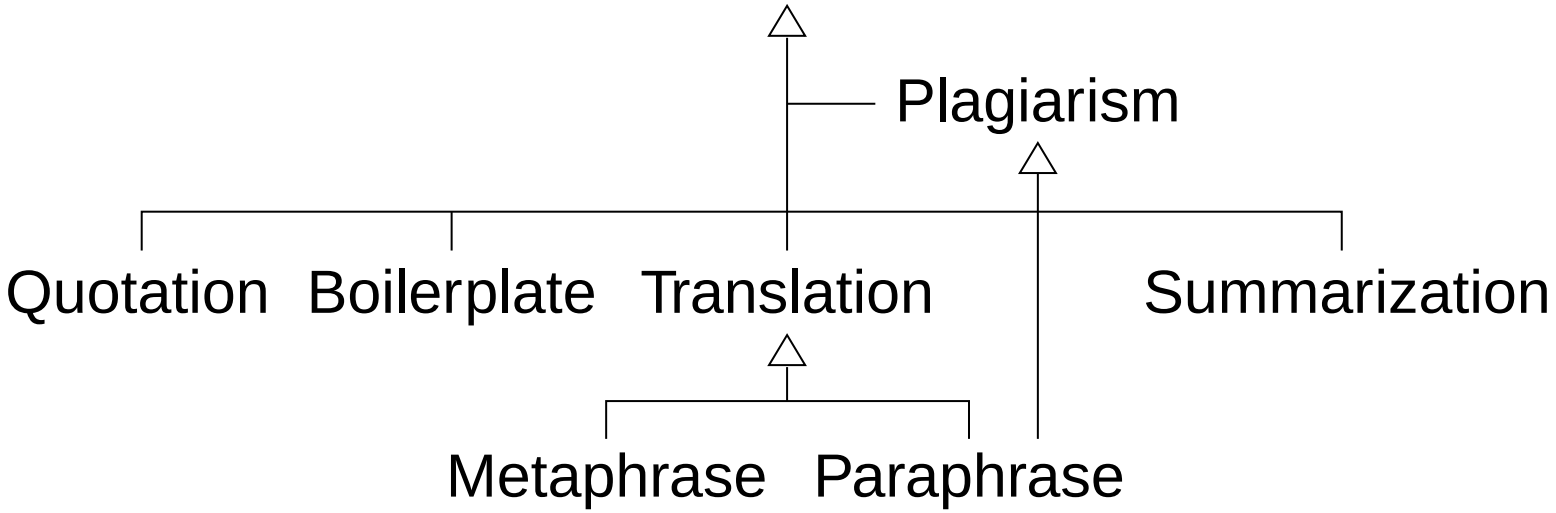
Technologies for Reusing Text from the Web

Technologies for Reusing Text from the Web

Technologies for Reusing Text from the Web



Technologies for Reusing Text from the Web



Contributions of Technologies for Reusing Text from the Web

1. Models & Algorithms
   - Unifying fingerprinting framework
   - Cross-language ESA
   - Comment cross-media similarity
   - Query segmentation algorithms

2. Surveys
   - Fingerprinting
   - Plagiarism detection
   - Web comment retrieval
   - Query segmentation

3. Evaluation Resources
   - Wikipedia as near-duplicate corpus
   - Wikipedia as cross-language corpus
   - 3 measures for plagiarism detection
   - 3 plagiarism corpora
   - Query segmentation corpus

4. Comparative Evaluations
   - 5 fingerprint algorithms
   - 3 cross-language models
   - 32 plagiarism detectors within 3 PAN evaluation competitions
   - 8 query segmentation algorithms

5. Tools
   - Netspeak
   - Picapica
   - OpinionCloud
   - Altools lib

Detecting Cross-Language Text Reuse

Measuring Cross-language Similarity

Alan Turing was conceived at Chatrapur, Orissa, India. His father was a member of the Indian Civil Service. He and his wife wanted Alan to be brought up in England, so they returned to Maida Vale, London, where Alan Turing was born on 23 June 1912. He had an elder brother, John. His father's civil service commission was still active, and during Turing's childhood years his parents travelled between Hastings, England and India, leaving their two sons to stay with a retired Army couple. Very early in life, Turing showed signs of the genius he was to later prominently display.

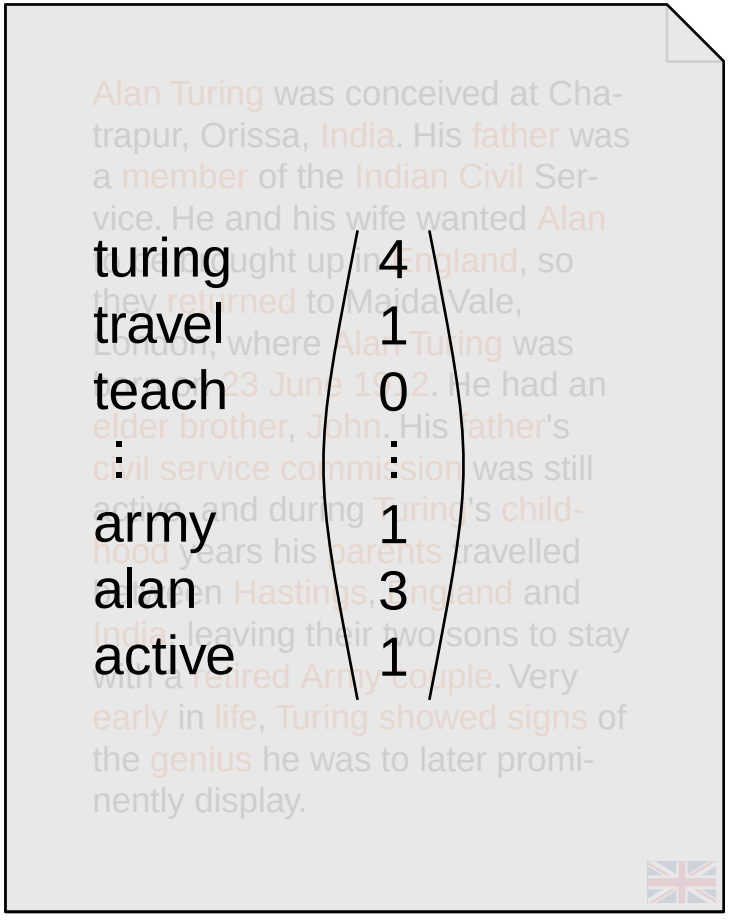
Alan Mathison Turing was born on 23 June 1912. His father was Julius Mathison Turing, member of the civil service in India, and his mother Ethel Sara Turing, the daughter of Edward Waller Stoney. Alan's childhood was spent with his elder brother John, living with a retired Army couple near Hastings, England. His parents returned to India until the end of his father‘s civil service commission, and visited when they could. Signs of Turing‘s genius showed early in his life. It is reported that he taught himself reading in less than three weeks.

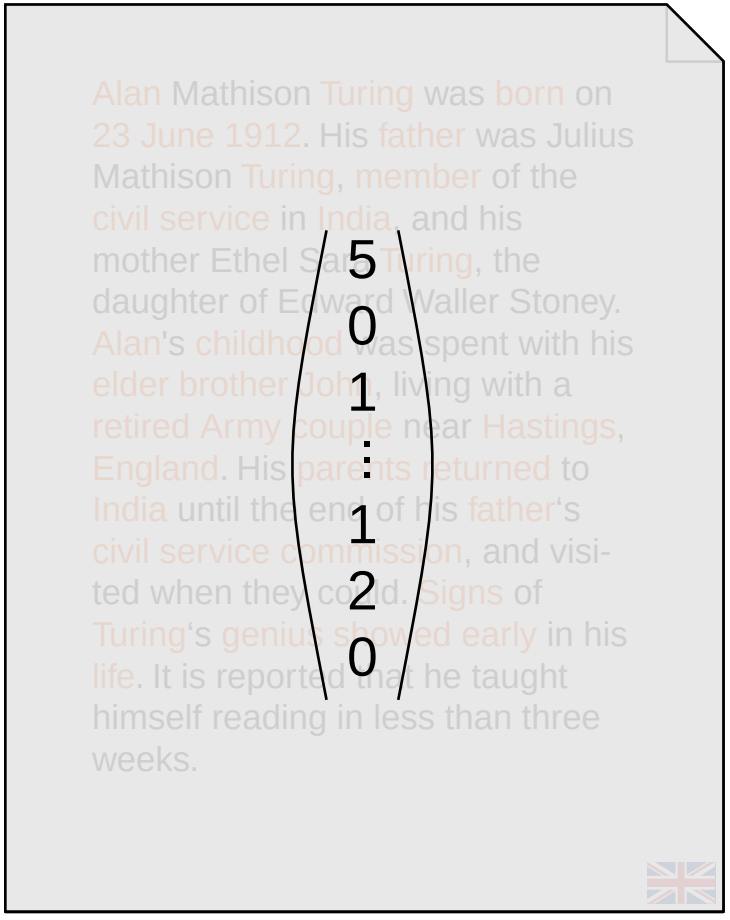
Measuring Cross-language Similarity

Alan Turing was conceived at Chatrapur, Orissa, India. His father was a member of the Indian Civil Service. He and his wife wanted Alan to be brought up in England, so they returned to Maida Vale, London, where Alan Turing was born on 23 June 1912. He had an elder brother, John. His father's civil service commission was still active, and during Turing's childhood years his parents travelled between Hastings, England and India, leaving their two sons to stay with a retired Army couple. Very early in life, Turing showed signs of the genius he was to later prominently display.

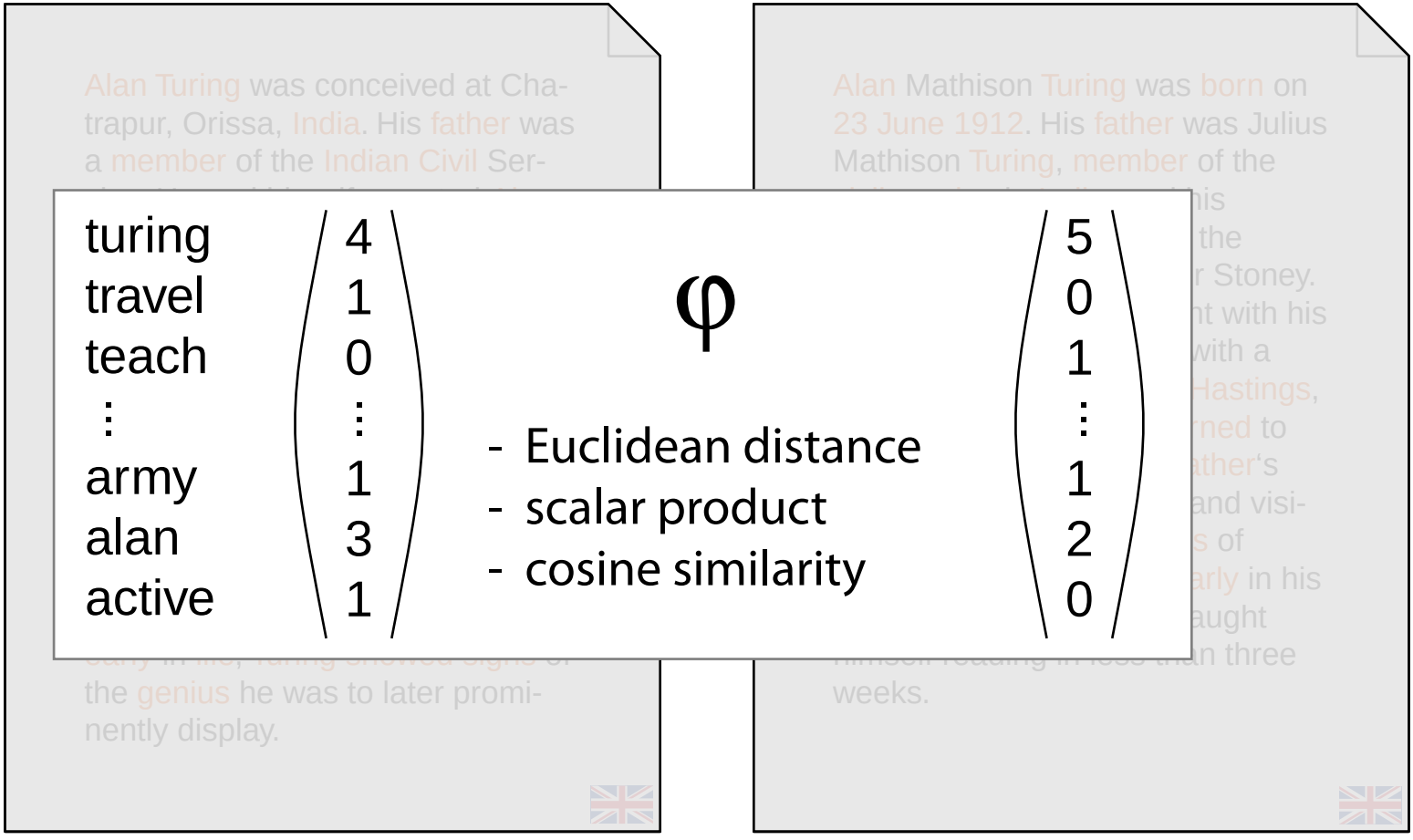
Alan Mathison Turing was born on 23 June 1912. His father was Julius Mathison Turing, member of the civil service in India, and his mother Ethel Sara Turing, the daughter of Edward Waller Stoney. Alan's childhood was spent with his elder brother John, living with a retired Army couple near Hastings, England. His parents returned to India until the end of his father‘s civil service commission, and visited when they could. Signs of Turing‘s genius showed early in his life. It is reported that he taught himself reading in less than three weeks.

Measuring Cross-language Similarity





Measuring Cross-language Similarity



Measuring Cross-language Similarity

Alan Turing was conceived at Chatrapur, Orissa, India. His father was a member of the Indian Civil Service. He and his wife wanted Alan to be brought up in England, so they returned to Maida Vale, London, where Alan Turing was born on 23 June 1912. He had an elder brother, John. His father's civil service commission was still active, and during Turing's childhood years his parents travelled between Hastings, England and India, leaving their two sons to stay with a retired Army couple. Very early in life, Turing showed signs of the genius he was to later prominently display.

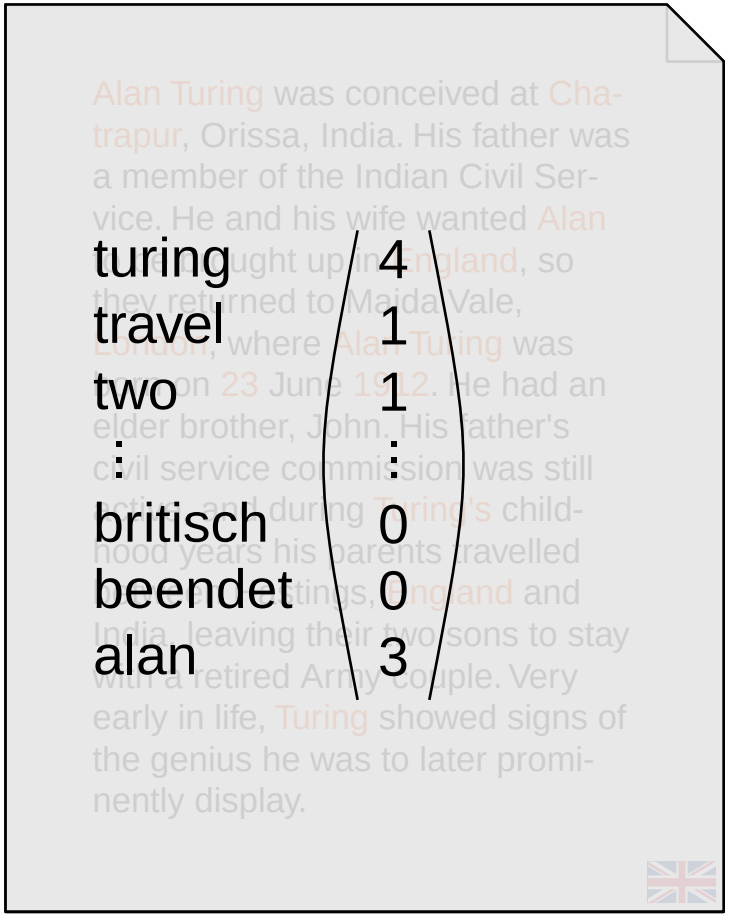
Turings Vater Julius Mathison Turing, ein britischer Staatsdiener in Chatrapur, Indien, und dessen Frau Ethel Sara wollten, dass ihr Kind in Großbritannien geboren wird. Deshalb kehrten sie nach London-Paddington zurück, wo Alan Turing am 23. Juni 1912 zur Welt kam. Da der Staatsdienst seines Vaters noch nicht beendet war, pendelte dieser während Turings Kindheit zwischen England und Indien. Seine Familie ließ er aus Furcht vor Gefahren in der britischen Kolonie bei Freunden in England zurück. Schon in frühester Kindheit zeigte sich die hohe Begabung und Intelligenz Turings.

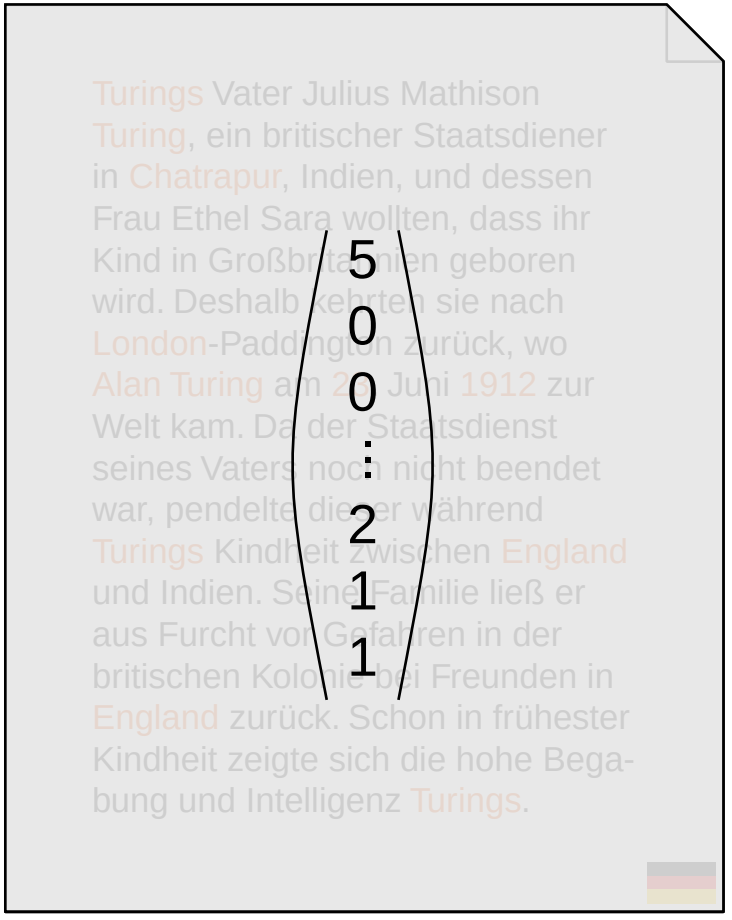
Measuring Cross-language Similarity

Alan Turing was conceived at Chatrapur, Orissa, India. His father was a member of the Indian Civil Service. He and his wife wanted Alan to be brought up in England, so they returned to Maida Vale, London, where Alan Turing was born on 23 June 1912. He had an elder brother, John. His father's civil service commission was still active, and during Turing's childhood years his parents travelled between Hastings, England and India, leaving their two sons to stay with a retired Army couple. Very early in life, Turing showed signs of the genius he was to later prominently display.

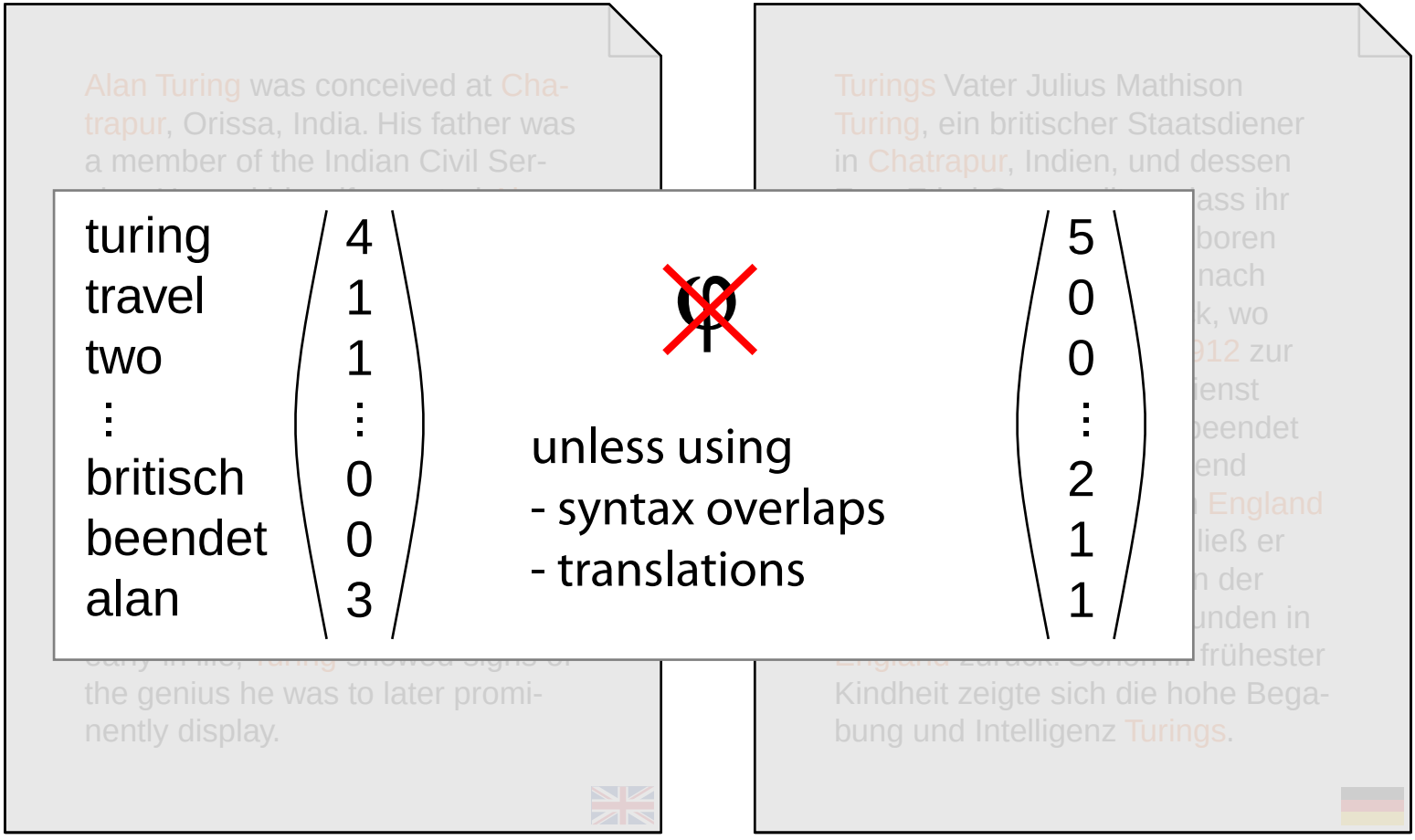
Turings Vater Julius Mathison Turing, ein britischer Staatsdiener in Chatrapur, Indien, und dessen Frau Ethel Sara wollten, dass ihr Kind in Großbritannien geboren wird. Deshalb kehrten sie nach London-Paddington zurück, wo Alan Turing am 23. Juni 1912 zur Welt kam. Da der Staatsdienst seines Vaters noch nicht beendet war, pendelte dieser während Turings Kindheit zwischen England und Indien. Seine Familie ließ er aus Furcht vor Gefahren in der britischen Kolonie bei Freunden in England zurück. Schon in frühester Kindheit zeigte sich die hohe Begabung und Intelligenz Turings.

Measuring Cross-language Similarity



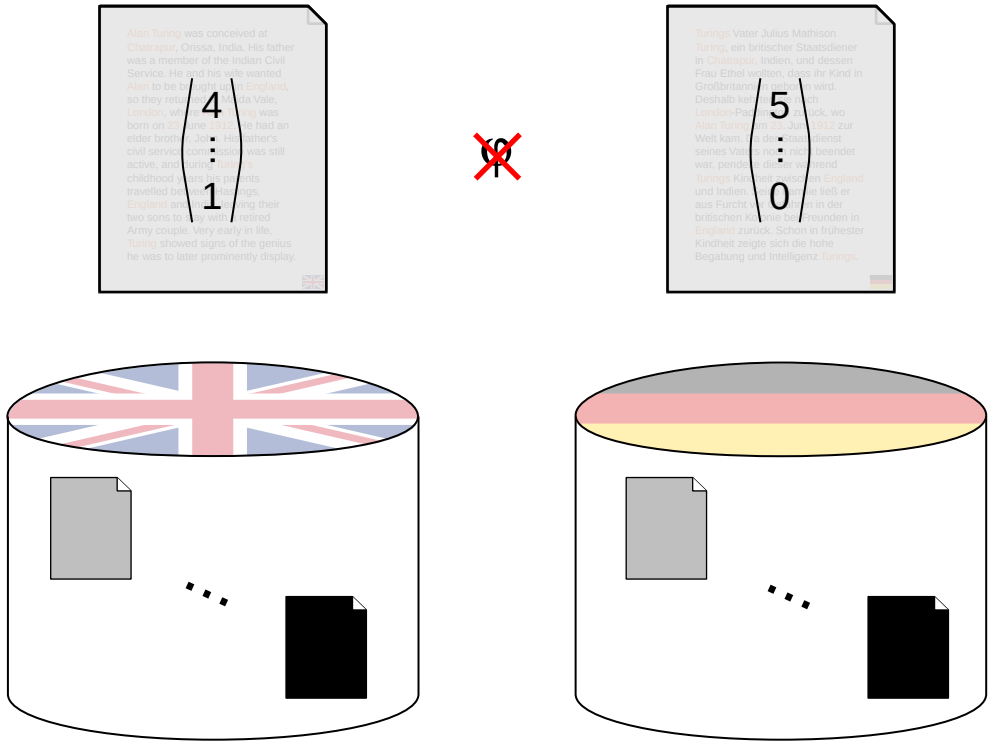


Measuring Cross-language Similarity

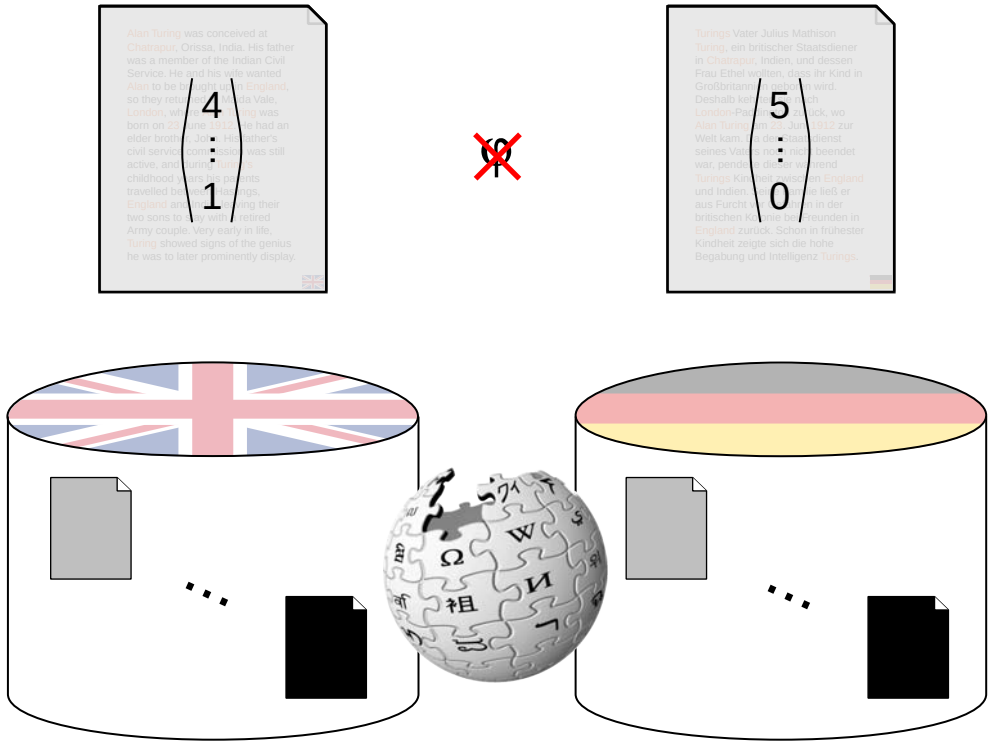


Cross-language Explicit Semantic Analysis

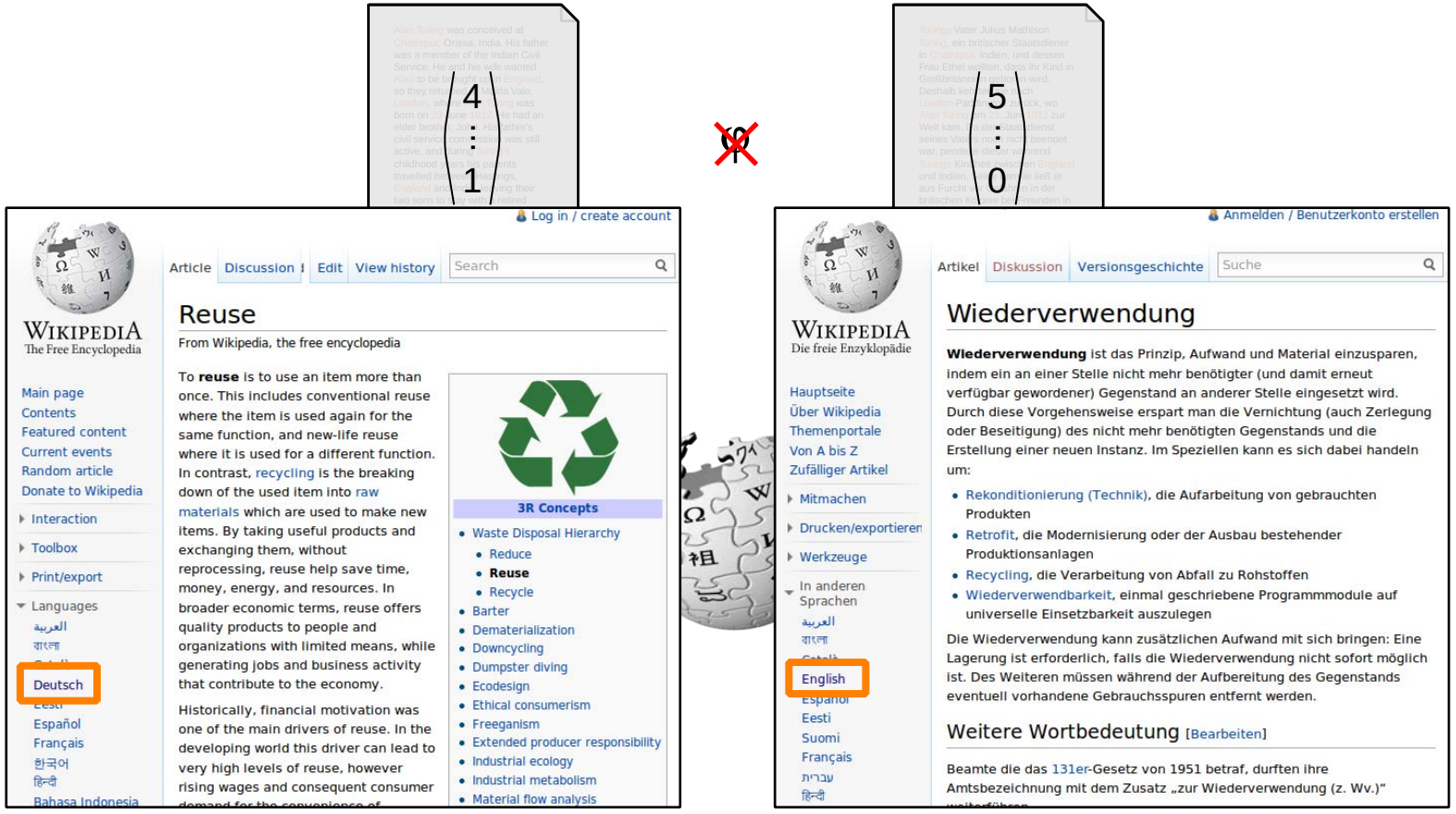
Cross-language Explicit Semantic Analysis



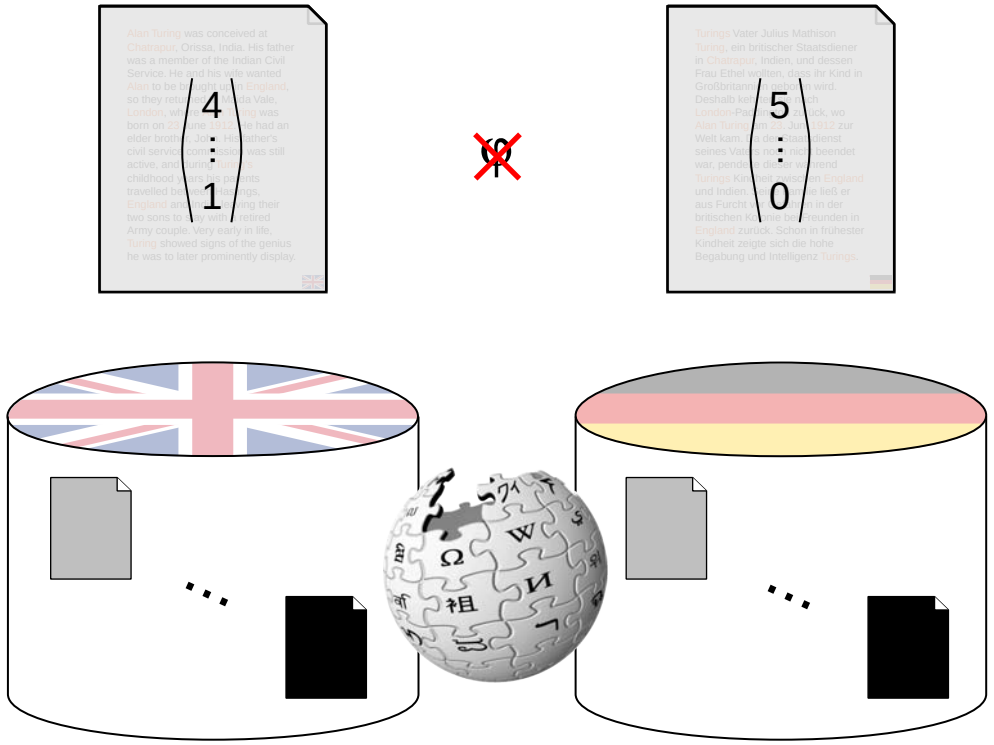
Cross-language Explicit Semantic Analysis



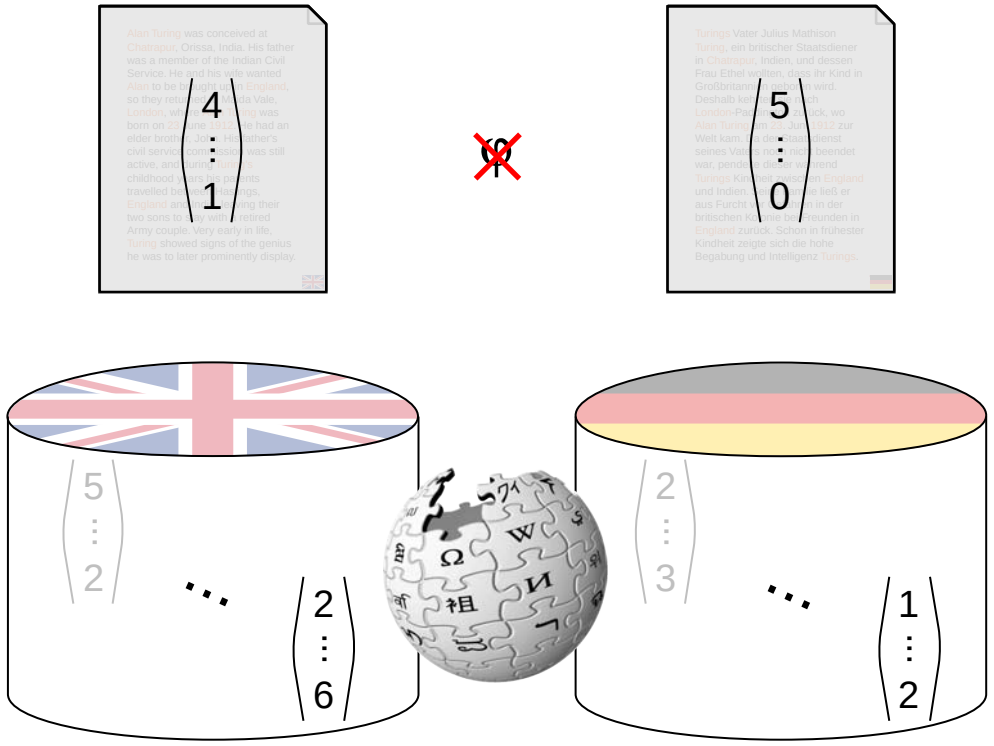
Cross-language Explicit Semantic Analysis



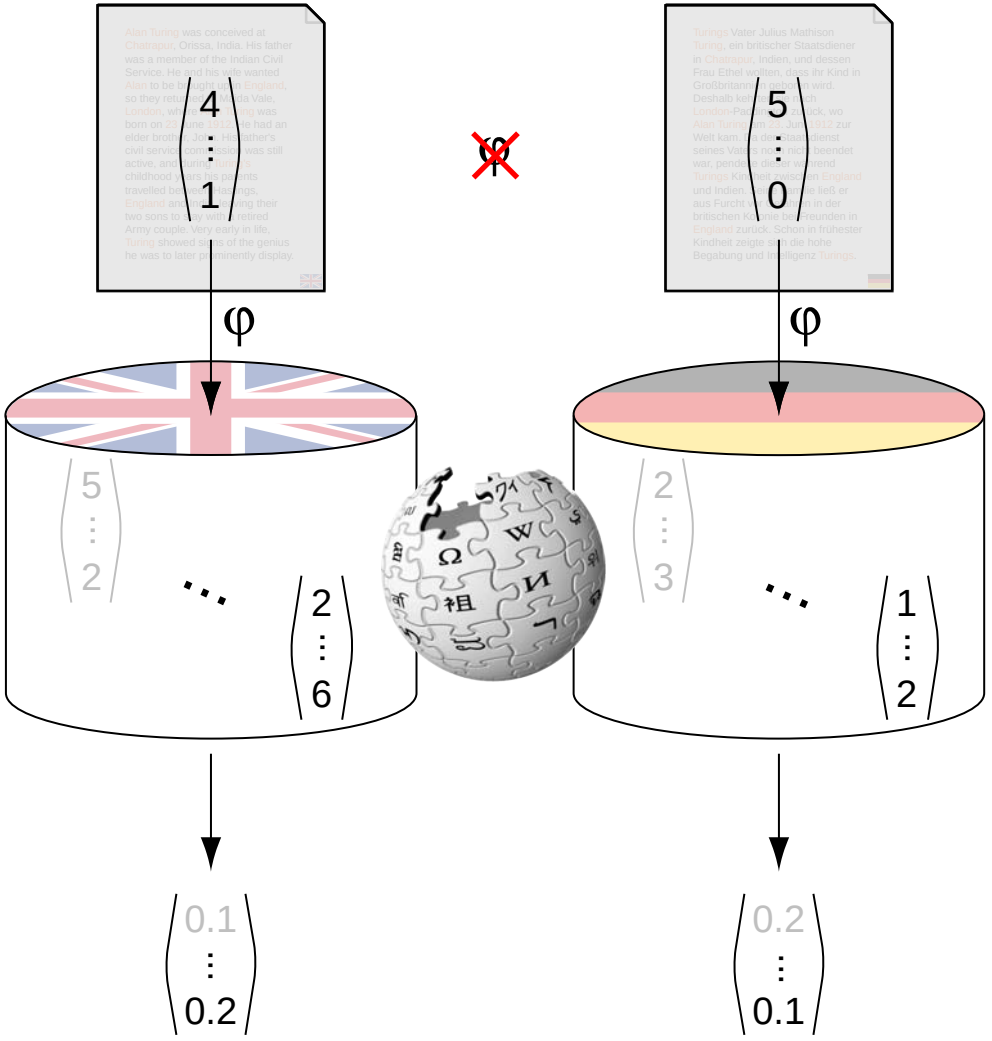
Cross-language Explicit Semantic Analysis



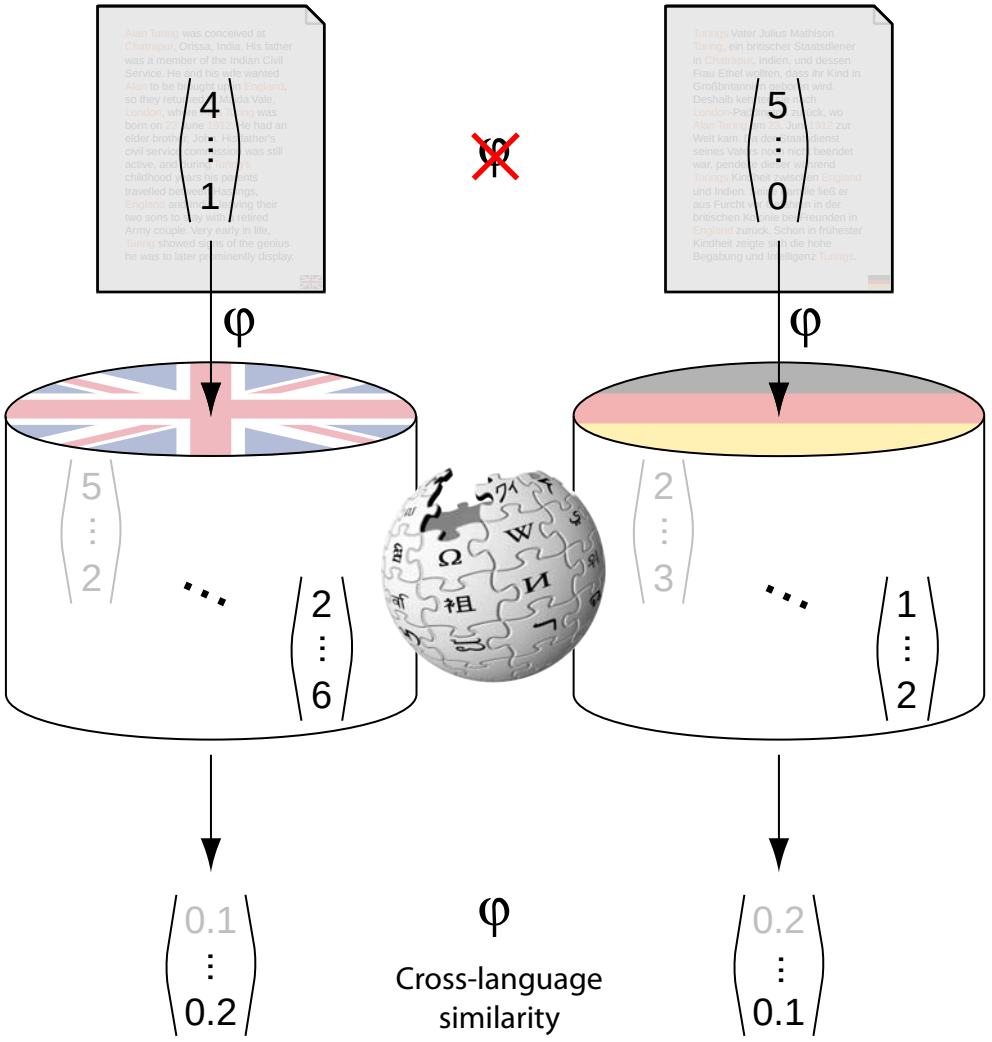
Cross-language Explicit Semantic Analysis



Cross-language Explicit Semantic Analysis



Cross-language Explicit Semantic Analysis



Cross-language Explicit Semantic Analysis

Experiments

1. cross-language ranking
2. bilingual rank correlation
3. cross-language similarity distribution
4. quality vs. dimensionality of CL-ESA
5. multilingualism (number of possible simultaneous languages)
6. runtime

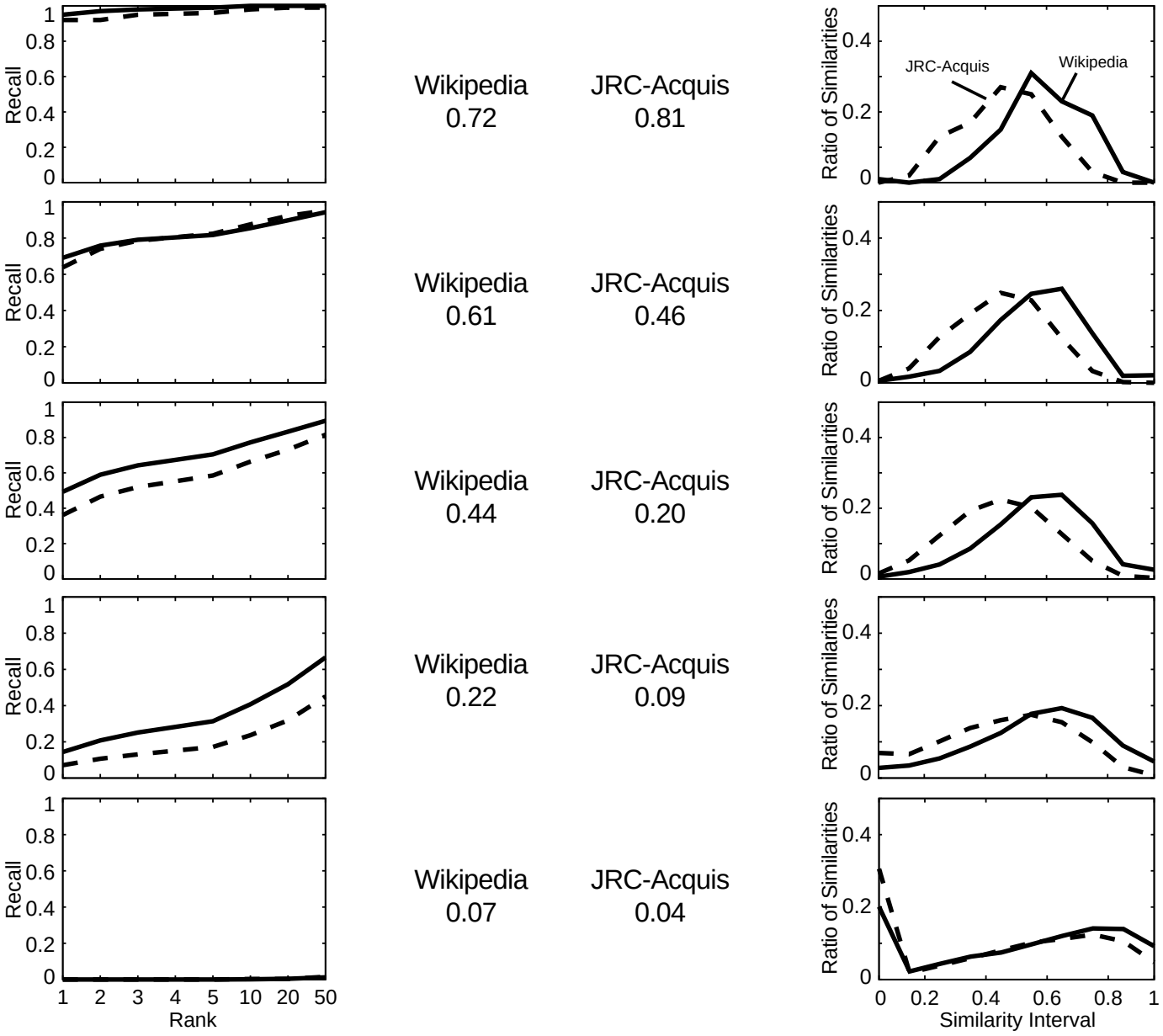
- comparison to two other state of the art models
- usage of 2 multilingual test collections
- comparison on 6 pairs of languages
- more than 100 000 documents in each of several dozen runs
- > 100 million similarities computed

Experiment 1
Cross-language Ranking

Experiment 2
Bilingual rank correlation

Experiment 3
Cross-language Similarity Distribution

Dimensions



Evaluating Plagiarism Detectors

Detection Performance Measures

Suspicious Document d_{plg}

Alan Mathison Turing, OBE, FRS (23 June 1912 – 7 June 1954), was an English mathematician, logician, cryptanalyst, and computer scientist. He was highly influential in the development of computer science, providing a formalisation of the concepts of "algorithm" and "computation" with the Turing machine, which played a significant role in the creation of the modern computer. Turing is widely considered to be the father of computer science and artificial intelligence. He was stockily built, had a high-pitched voice, and was talkative, witty, and somewhat donnish.

During the Second World War, Turing worked for the Government Code and Cypher School at Bletchley Park, Britain's codebreaking centre. For a time he was head of Hut 8, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German ciphers, including the method of the bombe, an electromechanical machine that could find settings for the Enigma machine.

After the war he worked at the National Physical Laboratory, where he created one of the first designs for the stored-program computer ACE. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

Turing's homosexuality resulted in a criminal prosecution in 1952, when homosexual acts were still illegal in the United Kingdom. He accepted treatment with female hormones (chemical castration) as an alternative to prison. He died in 1954, just over two weeks before his 42nd birthday, from cyanide poisoning. An inquest determined it was suicide; his mother and some others believed his death was accidental. On 10 September 2009, following an Internet campaign, British Prime Minister Gordon Brown made an official public apology on behalf of the British government for the way in which Turing was treated after the war.

Taken from http://en.wikipedia.org/wiki/Alan_Turing and post-edited to include material from the right hand text.

Source Document d_{src}

Alan Turing was born on 23 June, 1912, in London. His father was in the Indian Civil Service and Turing's parents lived in India until his father's retirement in 1926. Turing and his brother stayed with friends and relatives in England. Turing studied mathematics at Cambridge University, and subsequently taught there, working in the burgeoning world of quantum mechanics. It was at Cambridge that he developed the proof which states that automatic computation cannot solve all mathematical problems. This concept, also known as the Turing machine, is considered the basis for the modern theory of computation.

In 1936, Turing went to Princeton University in America, returning to England in 1938. He began to work secretly part-time for the British cryptanalytic department, the Government Code and Cypher School. On the outbreak of war he took up full-time work at its headquarters, Bletchley Park.

After the war, Turing turned his thoughts to the development of a machine that would logically process information. He worked first for the National Physical Laboratory (1945-1948). His plans were dismissed by his colleagues and the lab lost out on being the first to design a digital computer. It is thought that Turing's blueprint would have secured them the honour, as his machine was capable of computation speeds higher than the others. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

In 1952, Turing was arrested and tried for homosexuality, then a criminal offence. To avoid prison, he accepted injections of oestrogen for a year, which were intended to neutralise his libido. In that era, homosexuals were considered a security risk as they were open to blackmail. Turing's security clearance was withdrawn, meaning he could no longer work for GCHQ, the post-war successor to Bletchley Park.

He committed suicide on 7 June, 1954.

Taken from http://www.bbc.co.uk/history/people/alan_turing

Detection Performance Measures

Suspicious Document d_{plg}

Alan Mathison Turing, OBE, FRS (23 June 1912 – 7 June 1954), was an English mathematician, logician, cryptanalyst, and computer scientist. He was highly influential in the development of computer science, providing a formalisation of the concepts of "algorithm" and "computation" with the Turing machine, which played a significant role in the creation of the modern computer. Turing is widely considered to be the father of computer science and artificial intelligence. He was stockily built, had a high-pitched voice, and was talkative, witty, and somewhat donnish.

During the Second World War, Turing worked for the Government Code and Cypher School at Bletchley Park, Britain's codebreaking centre. For a time he was head of Hut 8, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German ciphers, including the method of the bombe, an electromechanical machine that could find settings for the Enigma machine.

s_{plg} After the war he worked at the National Physical Laboratory, where he created one of the first designs for the stored-program computer ACE. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

Turing's homosexuality resulted in a criminal prosecution in 1952, when homosexual acts were still illegal in the United Kingdom. He accepted treatment with female hormones (chemical castration) as an alternative to prison. He died in 1954, just over two weeks before his 42nd birthday, from cyanide poisoning. An inquest determined it was suicide; his mother and some others believed his death was accidental. On 10 September 2009, following an Internet campaign, British Prime Minister Gordon Brown made an official public apology on behalf of the British government for the way in which Turing was treated after the war.

Taken from http://en.wikipedia.org/wiki/Alan_Turing and post-edited to include material from the right hand text.

Source Document d_{src}

Alan Turing was born on 23 June, 1912, in London. His father was in the Indian Civil Service and Turing's parents lived in India until his father's retirement in 1926. Turing and his brother stayed with friends and relatives in England. Turing studied mathematics at Cambridge University, and subsequently taught there, working in the burgeoning world of quantum mechanics. It was at Cambridge that he developed the proof which states that automatic computation cannot solve all mathematical problems. This concept, also known as the Turing machine, is considered the basis for the modern theory of computation.

In 1936, Turing went to Princeton University in America, returning to England in 1938. He began to work secretly part-time for the British cryptanalytic department, the Government Code and Cypher School. On the outbreak of war he took up full-time work at its headquarters, Bletchley Park.

s_{src} After the war, Turing turned his thoughts to the development of a machine that would logically process information. He worked first for the National Physical Laboratory (1945-1948). His plans were dismissed by his colleagues and the lab lost out on being the first to design a digital computer. It is thought that Turing's blueprint would have secured them the honour, as his machine was capable of computation speeds higher than the others. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

In 1952, Turing was arrested and tried for homosexuality, then a criminal offence. To avoid prison, he accepted injections of oestrogen for a year, which were intended to neutralise his libido. In that era, homosexuals were considered a security risk as they were open to blackmail. Turing's security clearance was withdrawn, meaning he could no longer work for GCHQ, the post-war successor to Bletchley Park.

He committed suicide on 7 June, 1954.

Taken from http://www.bbc.co.uk/history/people/alan_turing

- Plagiarism $s = \langle s_{\text{plg}}, d_{\text{plg}}, s_{\text{src}}, d_{\text{src}} \rangle$

Detection Performance Measures

Suspicious Document d_{plg}

Alan Mathison Turing, OBE, FRS (23 June 1912 – 7 June 1954), was an English mathematician, logician, cryptanalyst, and computer scientist. He was highly influential in the development of computer science, providing a formalisation of the concepts of "algorithm" and "computation" with the Turing machine, which played a significant role in the creation of the modern computer. Turing is widely considered to be the father of computer science and artificial intelligence. He was stockily built, had a high-pitched voice, and was talkative, witty, and somewhat donnish.

During the Second World War, Turing worked for the Government Code and Cypher School at Bletchley Park, Britain's codebreaking centre. For a time he was head of Hut 8, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German ciphers, including the method of the bombe, an electromechanical machine that could find settings for the Enigma machine.

s_{plg} After the war he worked at the National Physical Laboratory, where he created one of the first designs for the stored-program computer ACE. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

Turing's homosexuality resulted in a criminal prosecution in 1952, when homosexual acts were still illegal in the United Kingdom. He accepted treatment with female hormones (chemical castration) as an alternative to prison. He died in 1954, just over two weeks before his 42nd birthday, from cyanide poisoning. An inquest determined it was suicide; his mother and some others believed his death was accidental. On 10 September 2009, following an Internet campaign, British Prime Minister Gordon Brown made an official public apology on behalf of the British government for the way in which Turing was treated after the war. r_{plg}

Taken from http://en.wikipedia.org/wiki/Alan_Turing and post-edited to include material from the right hand text.

Source Document d_{src}

Alan Turing was born on 23 June, 1912, in London. His father was in the Indian Civil Service and Turing's parents lived in India until his father's retirement in 1926. Turing and his brother stayed with friends and relatives in England. Turing studied mathematics at Cambridge University, and subsequently taught there, working in the burgeoning world of quantum mechanics. It was at Cambridge that he developed the proof which states that automatic computation cannot solve all mathematical problems. This concept, also known as the Turing machine, is considered the basis for the modern theory of computation.

In 1936, Turing went to Princeton University in America, returning to England in 1938. He began to work secretly part-time for the British cryptanalytic department, the Government Code and Cypher School. On the outbreak of war he took up full-time work at its headquarters, Bletchley Park.

s_{src} After the war, Turing turned his thoughts to the development of a machine that would logically process information. He worked first for the National Physical Laboratory (1945-1948). His plans were dismissed by his colleagues and the lab lost out on being the first to design a digital computer. It is thought that Turing's blueprint would have secured them the honour, as his machine was capable of computation speeds higher than the others. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

In 1952, Turing was arrested and tried for homosexuality, then a criminal offence. To avoid prison, he accepted injections of oestrogen for a year, which were intended to neutralise his libido. In that era, homosexuals were considered a security risk as they were open to blackmail. Turing's security clearance was withdrawn, meaning he could no longer work for GCHQ, the post-war successor to Bletchley Park.

He committed suicide on 7 June, 1954. r_{src}

Taken from http://www.bbc.co.uk/history/people/alan_turing

- Plagiarism $s = \langle s_{\text{plg}}, d_{\text{plg}}, s_{\text{src}}, d_{\text{src}} \rangle$
- Detection $r = \langle r_{\text{plg}}, d_{\text{plg}}, r_{\text{src}}, d'_{\text{src}} \rangle$

Detection Performance Measures

Suspicious Document d_{plg}

Alan Mathison Turing, OBE, FRS (23 June 1912 – 7 June 1954), was an English mathematician, logician, cryptanalyst, and computer scientist. He was highly influential in the development of computer science, providing a formalisation of the concepts of "algorithm" and "computation" with the Turing machine, which played a significant role in the creation of the modern computer. Turing is widely considered to be the father of computer science and artificial intelligence. He was stockily built, had a high-pitched voice, and was talkative, witty, and somewhat donnish.

During the Second World War, Turing worked for the Government Code and Cypher School at Bletchley Park, Britain's codebreaking centre. For a time he was head of Hut 8, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German ciphers, including the method of the bombe, an electromechanical machine that could find settings for the Enigma machine.

s_{plg} After the war he worked at the National Physical Laboratory, where he created one of the first designs for the stored-program computer ACE. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

Turing's homosexuality resulted in a criminal prosecution in 1952, when homosexual acts were still illegal in the United Kingdom. He accepted treatment with female hormones (chemical castration) as an alternative to prison. He died in 1954, just over two weeks before his 42nd birthday, from cyanide poisoning. An inquest determined it was suicide; his mother and some others believed his death was accidental. On 10 September 2009, following an Internet campaign, British Prime Minister Gordon Brown made an official public apology on behalf of the British government for the way in which Turing was treated after the war. r_{plg}

Taken from http://en.wikipedia.org/wiki/Alan_Turing and post-edited to include material from the right hand text.

Source Document d_{src}

Alan Turing was born on 23 June, 1912, in London. His father was in the Indian Civil Service and Turing's parents lived in India until his father's retirement in 1926. Turing and his brother stayed with friends and relatives in England. Turing studied mathematics at Cambridge University, and subsequently taught there, working in the burgeoning world of quantum mechanics. It was at Cambridge that he developed the proof which states that automatic computation cannot solve all mathematical problems. This concept, also known as the Turing machine, is considered the basis for the modern theory of computation.

In 1936, Turing went to Princeton University in America, returning to England in 1938. He began to work secretly part-time for the British cryptanalytic department, the Government Code and Cypher School. On the outbreak of war he took up full-time work at its headquarters, Bletchley Park.

s_{src} After the war, Turing turned his thoughts to the development of a machine that would logically process information. He worked first for the National Physical Laboratory (1945-1948). His plans were dismissed by his colleagues and the lab lost out on being the first to design a digital computer. It is thought that Turing's blueprint would have secured them the honour, as his machine was capable of computation speeds higher than the others. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

In 1952, Turing was arrested and tried for homosexuality, then a criminal offence. To avoid prison, he accepted injections of oestrogen for a year, which were intended to neutralise his libido. In that era, homosexuals were considered a security risk as they were open to blackmail. Turing's security clearance was withdrawn, meaning he could no longer work for GCHQ, the post-war successor to Bletchley Park.

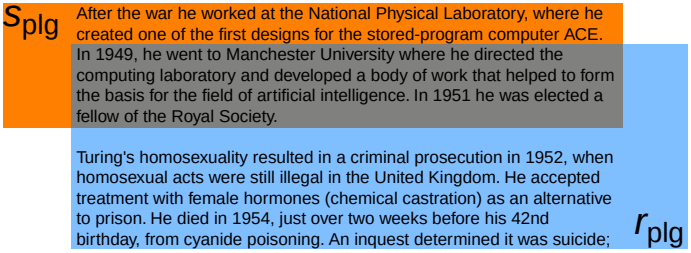
He committed suicide on 7 June, 1954. r_{src}

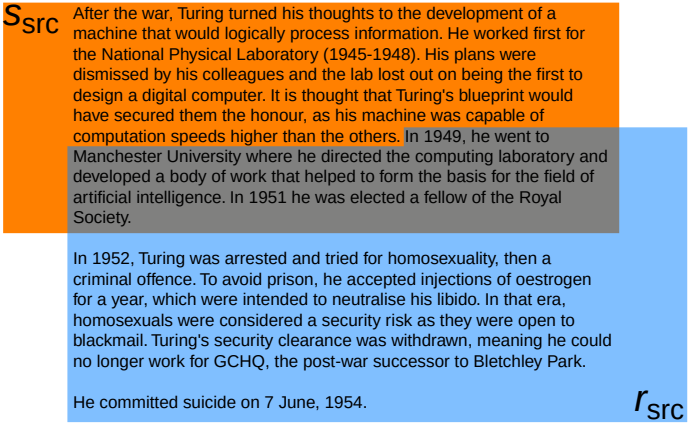
Taken from http://www.bbc.co.uk/history/people/alan_turing

- Plagiarism $s = \langle s_{\text{plg}}, d_{\text{plg}}, s_{\text{src}}, d_{\text{src}} \rangle$
- Detection $r = \langle r_{\text{plg}}, d_{\text{plg}}, r_{\text{src}}, d'_{\text{src}} \rangle$

- What is the detection quality?

Detection Performance Measures

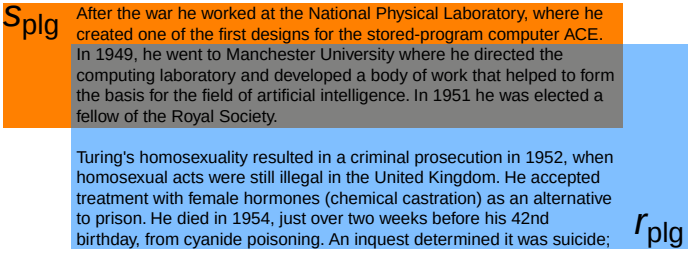


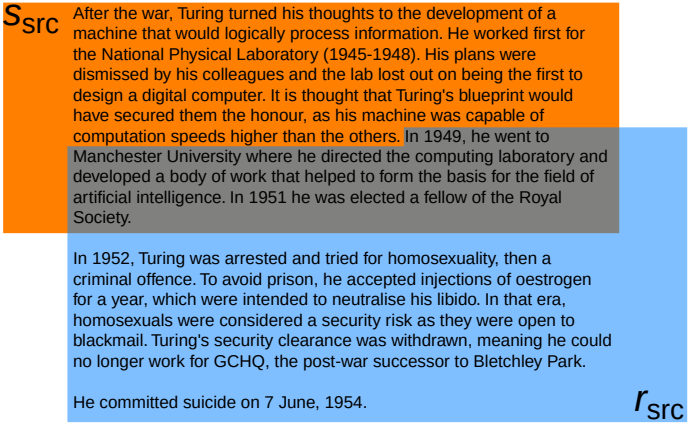


- Plagiarism $s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$
- Detection $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$

- What is the detection quality?

Detection Performance Measures



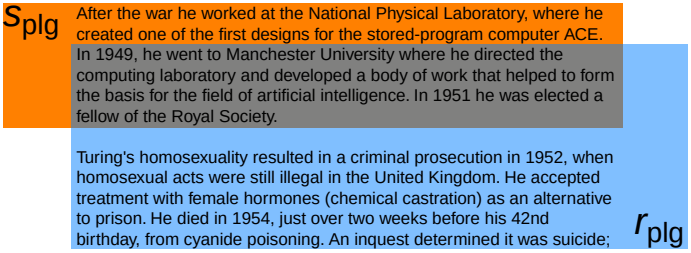


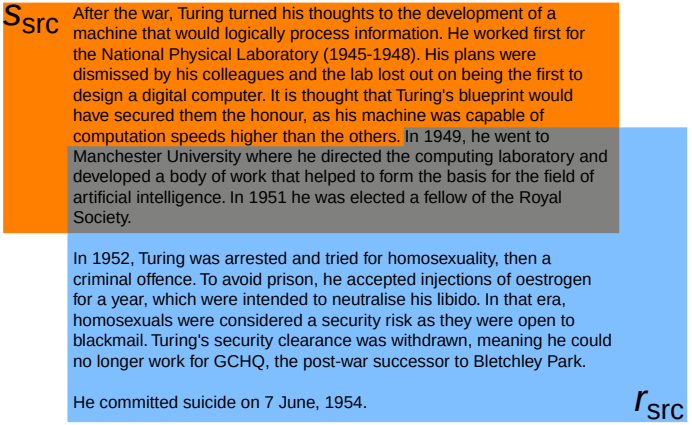
- Plagiarism $s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$
- Detection $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$

- What is the detection quality?

- r detects s iff $r_{plg} \cap s_{plg} \neq \emptyset$, $r_{src} \cap s_{src} \neq \emptyset$, and $d'_{src} = d_{src}$

Detection Performance Measures





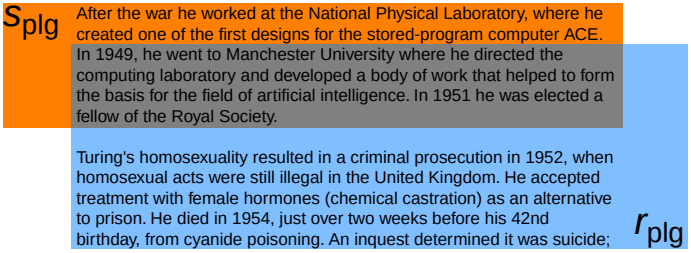
- Plagiarism $s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$
- Detection $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$

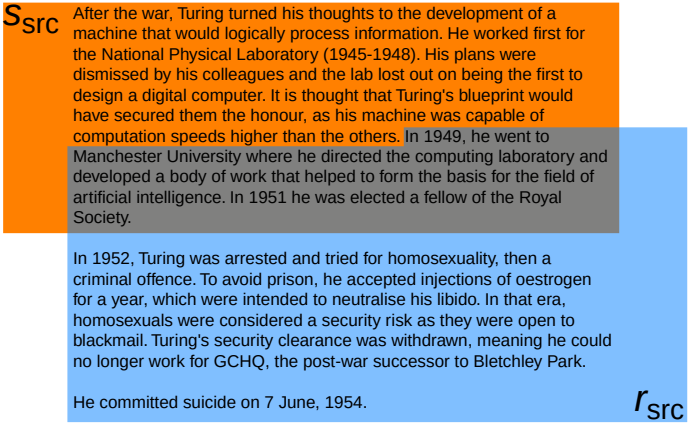
- What is the detection quality?

- r detects s iff $r_{plg} \cap s_{plg} \neq \emptyset$, $r_{src} \cap s_{src} \neq \emptyset$, and $d'_{src} = d_{src}$

- $|s \sqcap r| := \begin{cases} \text{number of overlapping characters} & \text{if } r \text{ detects } s, \\ 0 & \text{else} \end{cases}$

Detection Performance Measures





- Plagiarism $s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$
- Detection $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$

- What is the detection quality?

- r detects s iff $r_{plg} \cap s_{plg} \neq \emptyset$, $r_{src} \cap s_{src} \neq \emptyset$, and $d'_{src} = d_{src}$

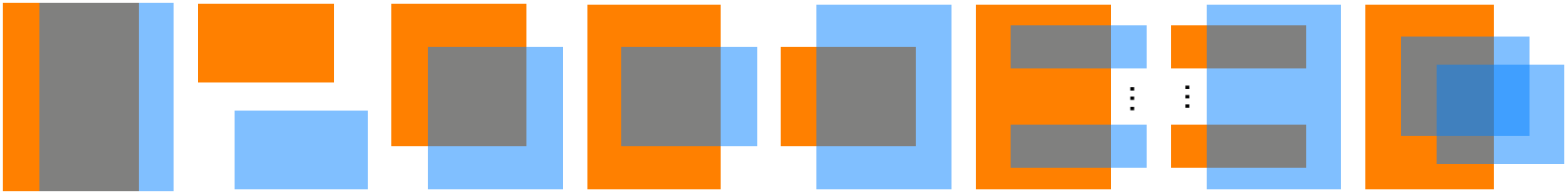
- $|s \sqcap r| := \begin{cases} \text{number of overlapping characters} & \text{if } r \text{ detects } s, \\ 0 & \text{else} \end{cases}$

- *precicion*$(s, r) = \dfrac{|s \sqcap r|}{|r|} = 0.38$

- *recall*$(s, r) = \dfrac{|s \sqcap r|}{|s|} = 0.45$

Detection Performance Measures

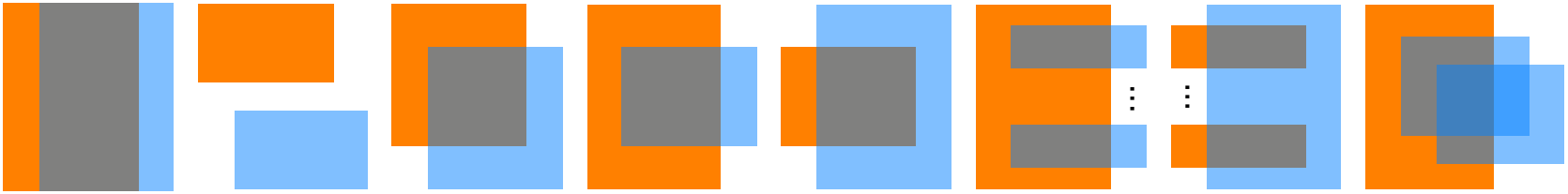
Possible patterns:



+ combinations thereof
+ combinations regarding pairs of suspicious and source documents

Detection Performance Measures

Possible patterns:

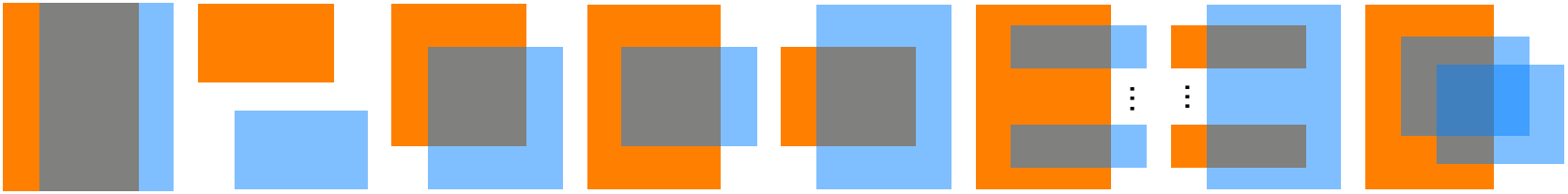


+ combinations thereof
+ combinations regarding pairs of suspicious and source documents

- no 1:1 correspondence between plagiarism cases and detections
- deal with sets of detections R and plagiarism cases S
- avoid double-counting of detection overlaps (inclusion-exclusion principle)

Detection Performance Measures

Possible patterns:



+ combinations thereof
+ combinations regarding pairs of suspicious and source documents

- no 1:1 correspondence between plagiarism cases and detections
- deal with sets of detections R and plagiarism cases S
- avoid double-counting of detection overlaps (inclusion-exclusion principle)
- measure precision for each detection and recall for each plagiarism case, averaging the results:

$$\mathit{precicion}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{\left| \bigcup_{s \in S} (\mathbf{s} \sqcap \mathbf{r}) \right|}{|\mathbf{r}|}$$

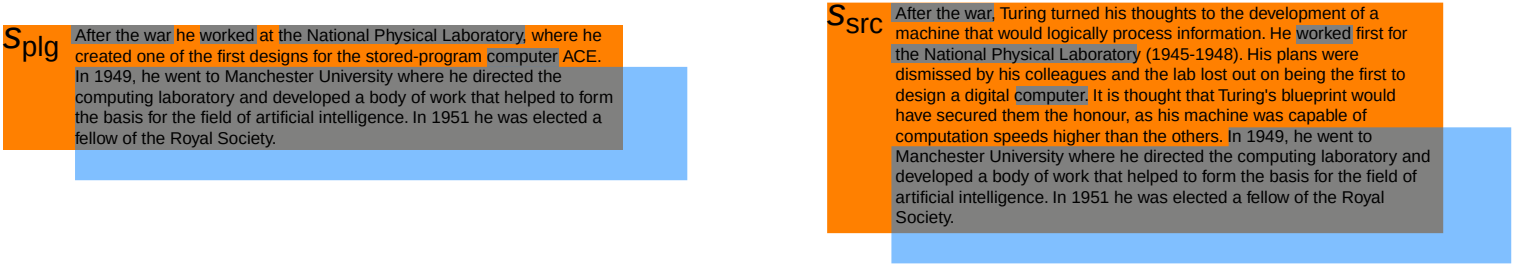
$$\mathit{recall}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{\left| \bigcup_{r \in R} (\mathbf{s} \sqcap \mathbf{r}) \right|}{|\mathbf{s}|}$$

Detection Performance Measures

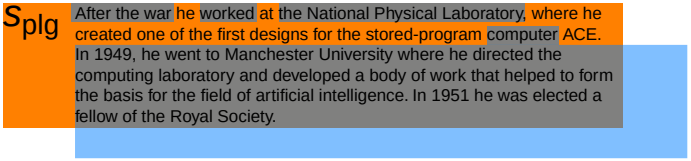
s_{plg} After the war he worked at the National Physical Laboratory, where he created one of the first designs for the stored-program computer ACE. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

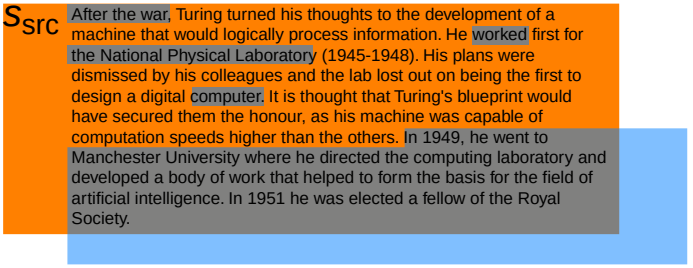
s_{src} After the war, Turing turned his thoughts to the development of a machine that would logically process information. He worked first for the National Physical Laboratory (1945-1948). His plans were dismissed by his colleagues and the lab lost out on being the first to design a digital computer. It is thought that Turing's blueprint would have secured them the honour, as his machine was capable of computation speeds higher than the others. In 1949, he went to Manchester University where he directed the computing laboratory and developed a body of work that helped to form the basis for the field of artificial intelligence. In 1951 he was elected a fellow of the Royal Society.

Detection Performance Measures



Detection Performance Measures



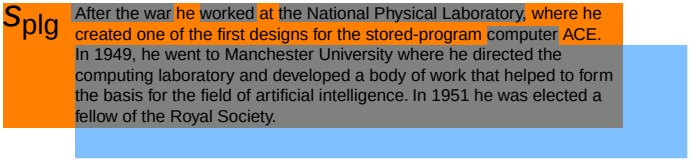


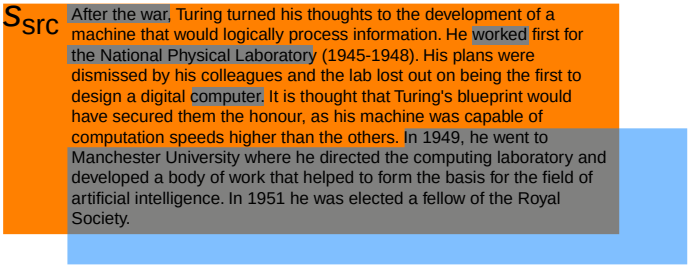
- undesirable fragmentation of the detection
- measure the average number of times a plagiarism case is detected:

$$\mathit{granularity}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|$$

where $S_R \subseteq S$ are detected cases, and $R_s \subseteq R$ are detections of s

Detection Performance Measures





- undesirable fragmentation of the detection
- measure the average number of times a plagiarism case is detected:

$$\mathit{granularity}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|$$

where $S_R \subseteq S$ are detected cases, and $R_s \subseteq R$ are detections of s

- *precicion*, *recall*, and *granularity* allow only for a partial order
- combination of the three measures into one score:

$$\mathit{plagdet}(S, R) = \frac{F_1}{\log_2(1 + \mathit{granularity}(S, R))}$$

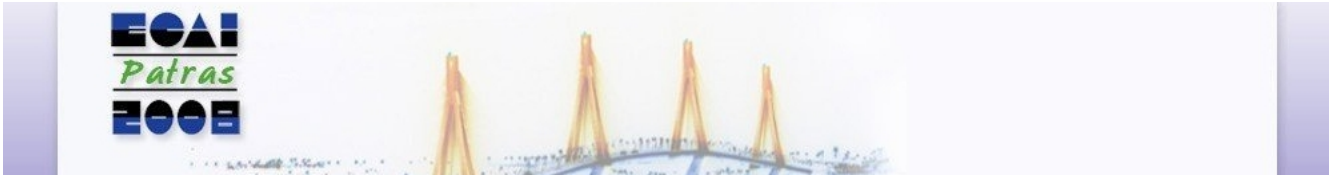
where F_1 is the harmonic mean of *precicion* and *recall*

Evaluation Competitions at PAN 2009-2011

Evaluation Competitions at PAN 2009-2011

2007

2008

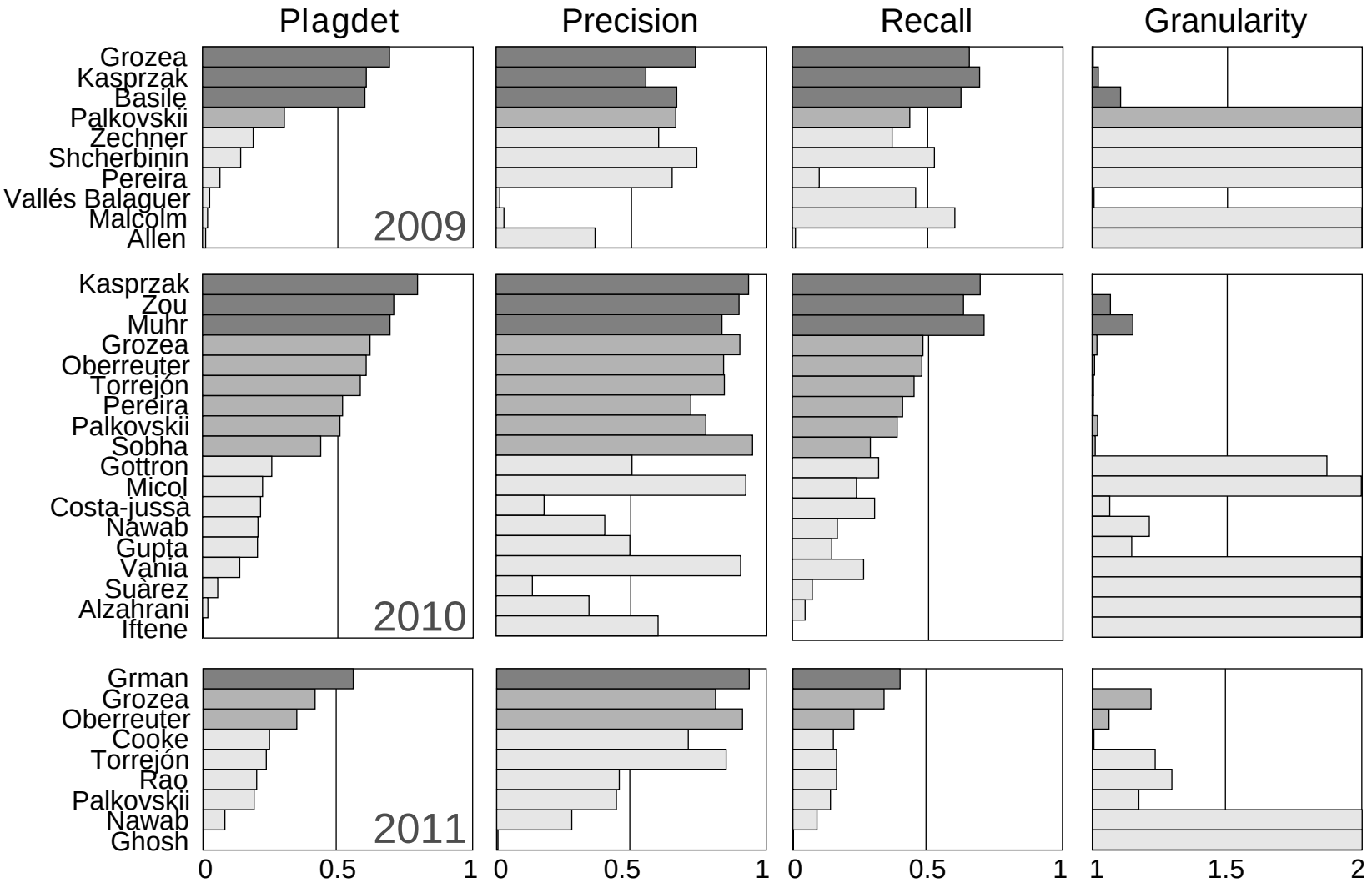


2009

2010

2011

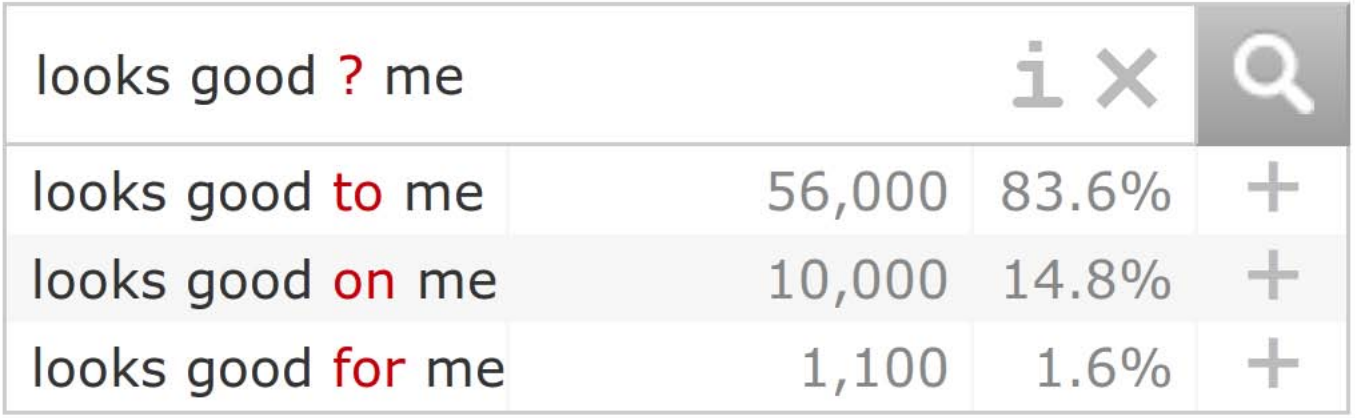
Evaluation Competitions at PAN 2009-2011



Reusing the Web for Writing Assistance

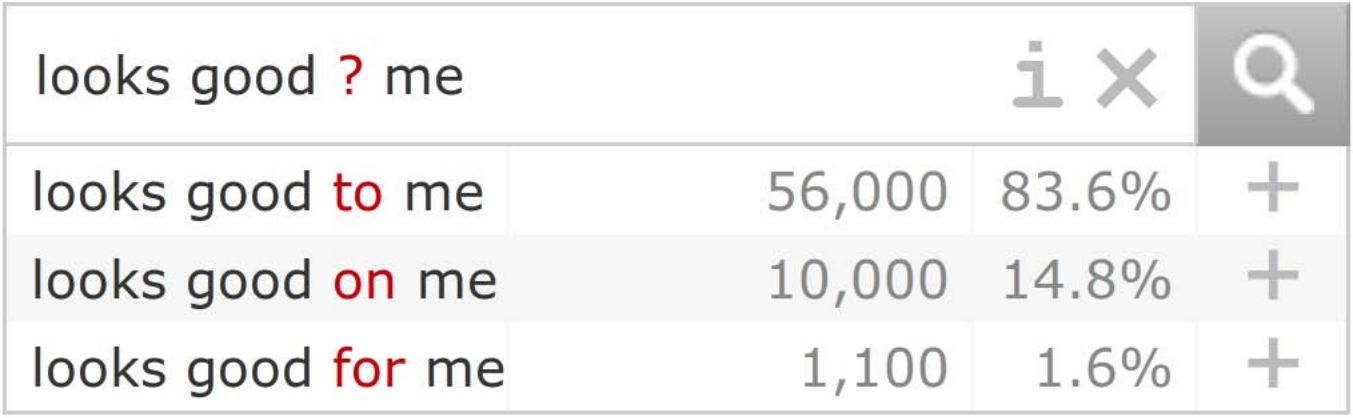
Reusing the Web for Writing Assistance

- writing is not so much about what to write, but how
- finding the right words is essential to maximize understanding
- Netspeak is a search engine for words in context:



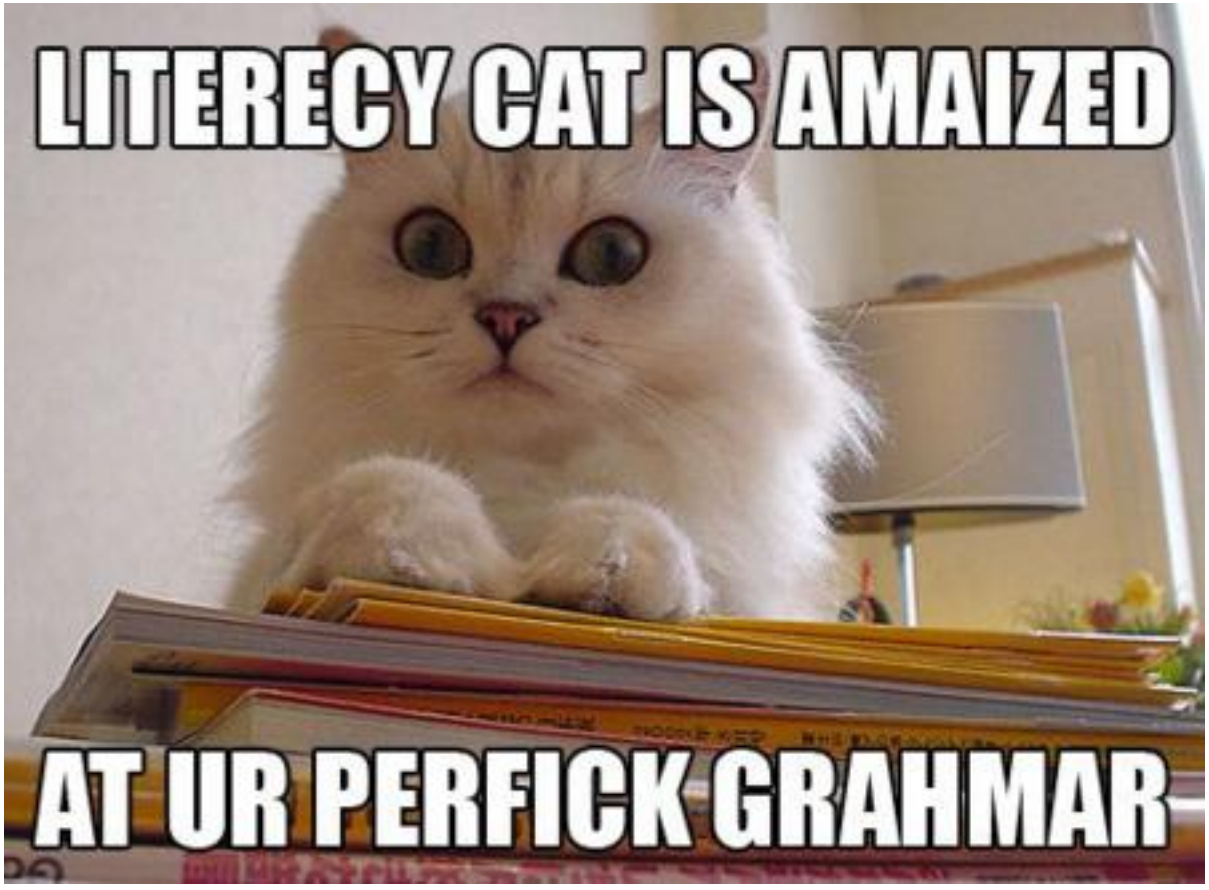
Reusing the Web for Writing Assistance

- writing is not so much about what to write, but how
- finding the right words is essential to maximize understanding
- Netspeak is a search engine for words in context:



Technical details:

- > 3 billion phrases and their usage frequencies as of 2006.
- > 120 GB inverted index data structure (scalable)
- < 1 second response time
- > 4300 users / month
- wildcard query processor
- instant search



Contributions of Technologies for Reusing Text from the Web

1. Models & Algorithms
   - Unifying fingerprinting framework
   - Cross-language ESA
   - Comment cross-media similarity
   - Query segmentation algorithms

2. Surveys
   - Fingerprinting
   - Plagiarism detection
   - Web comment retrieval
   - Query segmentation

3. Evaluation Resources
   - Wikipedia as near-duplicate corpus
   - Wikipedia as cross-language corpus
   - 3 measures for plagiarism detection
   - 3 plagiarism corpora
   - Query segmentation corpus

4. Comparative Evaluations
   - 5 fingerprint algorithms
   - 3 cross-language models
   - 32 plagiarism detectors within 3 PAN evaluation competitions
   - 8 query segmentation algorithms

5. Tools
   - Netspeak
   - Picapica
   - OpinionCloud
   - Altools lib

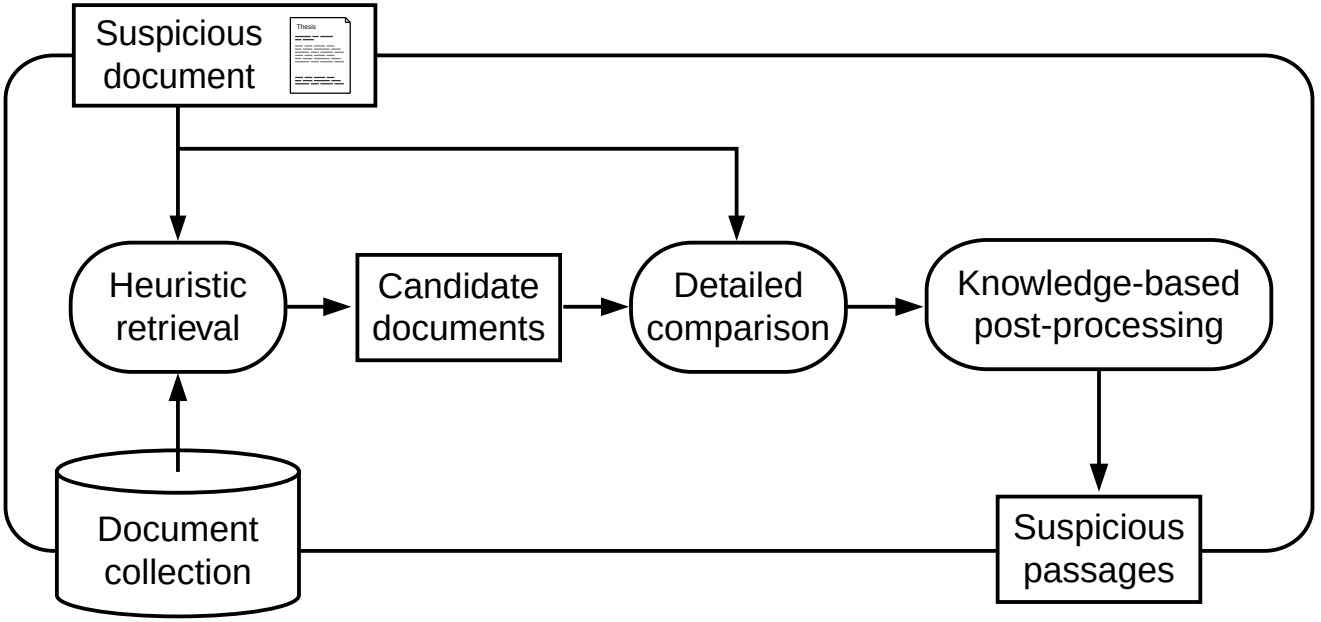
Benno Stein ❑ Maik Anderka ❑ Steven Burrows ❑ Tim Gollub ❑ Matthias Hagen ❑ Dennis Hoppe ❑ Nedim Lipka ❑ Sven Meyer zu Eißen ❑ Peter Prettenhofer ❑ Patrick Riehmann ❑ Bernd Fröhlich ❑ Alberto Barrón-Cedeño ❑ Paolo Rosso ❑ Paul Clough ❑ Steffen Becker ❑ Christof Bräutigam ❑ Andreas Eiselt ❑ Robert Gerling ❑ Teresa Holfeld ❑ Alexander Kümmel ❑ Fabian Loose ❑ Martin Trenkmann ❑ Dietmar Bratke ❑ Jürgen Eismann ❑ Nadin Glaser ❑ Maria-Theresa Hansens ❑ Melanie Hennig ❑ Dana Horch ❑ Antje Klahn ❑ Hildegard Kühndorf ❑ Tina Meinhardt ❑ Christin Oehmichen ❑ Ursula Schmidt ❑ Katja Schöllner ❑ Nils Rethmeier ❑ Tsvetomira Palakarska ❑ Steven Rönisch ❑ Hagen Christian Tönnies ❑ Michael Völske ❑ Anita Schilling ❑ Michael Bierstedt ❑ Christoph Lössnitz ❑ Dennis Braunsdorf ❑ Alexander Kleppe ❑ Franz Coriand ❑ Verena Skuk ❑ Anne Köpsel ❑ Marcel Heunemann ❑ Stefan Knoblauch ❑ Klaus Krämer ❑ Christian Fricke ❑ Denis Kreis ❑ Clement Welsch ❑ Maximilian Michel ❑ Jan Grassegger ❑ Jan Dittrich ❑ Fabian Vogelsteller ❑ Felicitas Höbelt ❑ Carsten Tetens ❑ Jan Hühne ❑ Nils Gründl ❑ André Zölitz ❑ Michael Hengst ❑ Yunlu Ai ❑ Markus Riedel ❑ Bjarne-Vanja Melani ❑ Henning Gründl ❑ Stephan Bongartz ❑ Daniel, Wiebke, Marc und Merle Potthast ❑ Steffi, Leonie und Louisa Daniel ❑ Gabi und Günter Aab ❑ Georg Potthast und Hildegard Knoke ❑ Ellinor Pfützner ❑ Martin Weitert ❑ Daniel Warner ❑ Christian Ederer

Thank you!

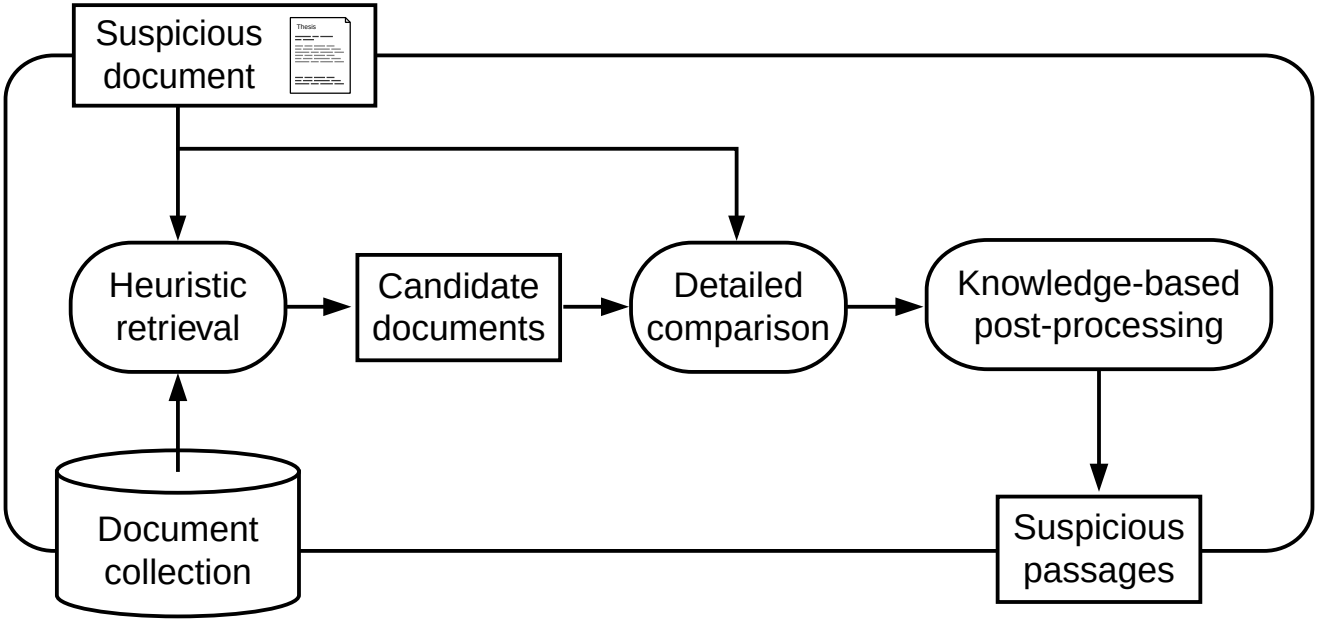
Appendix

- Detecting Plagiarism and Evaluating Detectors
- Survey of Plagiarism Detection Evaluations
- Plagiarism Corpus Construction
- Netspeak Experiments

Detecting Plagiarism



Detecting Plagiarism



Evaluating Plagiarism Detectors

Simulate inputs — measure output quality — repeat

What's required:

- corpus of plagiarism cases
- performance mesaures
- alternative implementations

Survey of Plagiarism Detection Evaluations

Evaluation Aspect	Text	Code
Experiment Task		
local collection	80%	95%
Web retrieval	15%	0%
other	5%	5%
Performance Measure		
precision, recall	43%	18%
manual, similarity	35%	69%
runtime only	15%	1%
other	7%	12%
Comparison		
none	46%	51%
parameter settings	19%	9%
other algorithms	35%	40%

Evaluation Aspect	Text	Code
Corpus Acquisition		
existing corpus	20%	18%
homemade corpus	80%	82%
Corpus Size [# documents]		
$[1, 10)$	11%	10%
$[10, 10^2)$	19%	30%
$[10^2, 10^3)$	38%	33%
$[10^3, 10^4)$	8%	11%
$[10^4, 10^5)$	16%	4%
$[10^5, 10^6)$	8%	0%

- more than 200 papers were reviewed
- many struggle with proper evaluation

Plagiarism Corpus Construction

Corpus overview:

- real plagiarism cases not available on a large scale
- plagiarism was generated automatically using heuristics
- plagiarism was also crowdsourced via Amazon's Mechanical Turk
- the corpus was compiled 3 years in a row, improving it each time
- $\sim$ 27 000 documents (obtained from the Project Gutenberg)
- $\sim$ 61 000 plagiarism cases

Plagiarism Corpus Construction

Corpus overview:

- real plagiarism cases not available on a large scale
- plagiarism was generated automatically using heuristics
- plagiarism was also crowdsourced via Amazon's Mechanical Turk
- the corpus was compiled 3 years in a row, improving it each time
- $\sim$ 27 000 documents (obtained from the Project Gutenberg)
- $\sim$ 61 000 plagiarism cases

Corpus parameters:

1. document length
2. document purpose
3. plagiarism per document
4. plagiarism case length
5. plagiarism case obfuscation

Corpus Parameters

100%	26 939 documents

Corpus Parameters

100% 26 939 documents

Document length:

50% 1-10 pages	35% 10-100 pages	15% 10^2-10^3 pp.

Document purpose:

50% source documents	50% suspicious documents

Plagiarism per suspicious document:

50% none	50% range from little to entirely

Corpus Parameters

100% 26 939 documents

Document length:

50% 1-10 pages	35% 10-100 pages	15% 10^2-10^3 pp.

Document purpose:

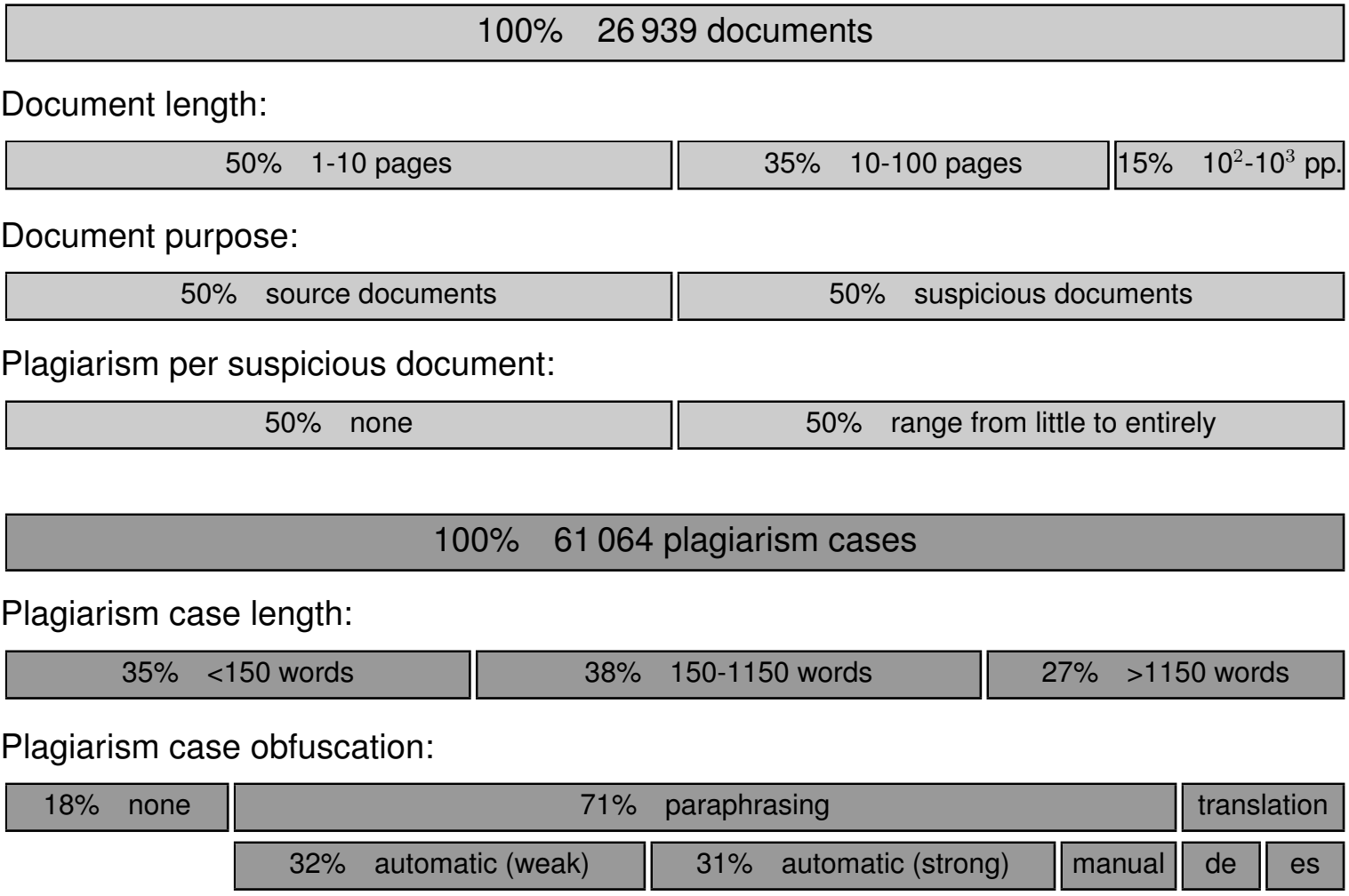
50% source documents	50% suspicious documents

Plagiarism per suspicious document:

50% none	50% range from little to entirely

100% 61 064 plagiarism cases

Corpus Parameters



- Manual paraphrases (8%) via Amazon's Mechanical Turk.
- Translations (11%) via Google Translate from de→en and es→en.

Netspeak Experiments