MORDOR : Myriadisation et Orchestration de Ramifications Divergentes pour l'Optimisation des Ressources

Marceau HERNANDEZ
Master II Sciences Du Langage
15/02/2024

Encadrant:
Gaël LEJEUNE

Introduction

Détour canonique: MazETTE

Merge without conflict ? Stratégies de fusion
- ROVER
- diff

Et maintenant ?

Introduction

Ressources textuelles et versions

Ressources textuelles Chaînes de caractères de longueur et alphabet variable.

- Deux ressources différentes peuvent être apparentées et n'être que deux versions d'un même *document*.
- Une version peut être la source d'une ou plusieurs autres.
- Plusieurs phénomènes peuvent être à l'origine de cette divergence, nous en considérerons deux familles

L'origine des versions: Ma version

- Plusieurs phénomènes peuvent être à l'origine de cette divergence, nous en considérerons deux familles:

Le changement de média Passer d'une ressource orale à textuelle, d'une image, d'une page web ou autre à un fichier texte, et cætera

Correction Se base sur le média original et la ressource textuelle

L'édition Modification volontaire, création d'une nouvelle version en se basant sur une ou plusieurs versions

L'origine des versions: Ma version

- Plusieurs phénomènes peuvent être à l'origine de cette divergence, nous en considérerons deux familles:

  Le changement de média Passer d'une ressource orale à textuelle, d'une image, d'une page web ou autre à un fichier texte, et cætera

  Correction Se base sur le média original et la ressource textuelle

  L'édition Modification volontaire, création d'une nouvelle version en se basant sur une ou plusieurs versions

1 version = 1 ou plusieurs de ces phénomènes

Exemple : un chanteur fait un *edit* de sa chanson et les paroles sont retranscrites sur plusieurs sites différents qui sont par la suite *scrappés* de manière différente

Une version pour les gouverner toutes ? Cas du changement de média

Parmi l'ensemble de ces versions, laquelle prendre ?

Une version pour les gouverner toutes ? Cas du changement de média

Parmi l'ensemble de ces versions, laquelle prendre ?

La meilleure version ne l'est pas forcément partout et le meilleur outil peut varier en performance selon le cas de figure

Exemple en OCR: page sale, sombre, avec des enluminures, etc.

Une version pour les gouverner toutes ? Cas du changement de média

Parmi l'ensemble de ces versions, laquelle prendre ?

La meilleure version ne l'est pas forcément partout et le meilleur outil peut varier en performance selon le cas de figure

Exemple en OCR: page sale, sombre, avec des enluminures, etc.

Intuition: Si 3/4 modèles donnent une réponse, c'est surement la bonne, même si le 1/4 est le plus performant en moyenne

Une version pour les gouverner toutes ? Cas du changement de média

Parmi l'ensemble de ces versions, laquelle prendre ?

La meilleure version ne l'est pas forcément partout et le meilleur outil peut varier en performance selon le cas de figure

Exemple en OCR: page sale, sombre, avec des enluminures, etc.

Intuition: Si 3/4 modèles donnent une réponse, c'est surement la bonne, même si le 1/4 est le plus performant en moyenne

Il nous faut donc **fusionner** les versions

Détour canonique: MazETTE

Mazette !

Mazarinades Extended Through Text Edition (MazETTE)

Plateforme de correction participative de retranscriptions issues d'OCR via myriadisation des annotateurs fondée autour de l'idée de fusion de versions.

Disponible au lien suivant :
`https://mazette.marceau-h.fr/`

Mazette !

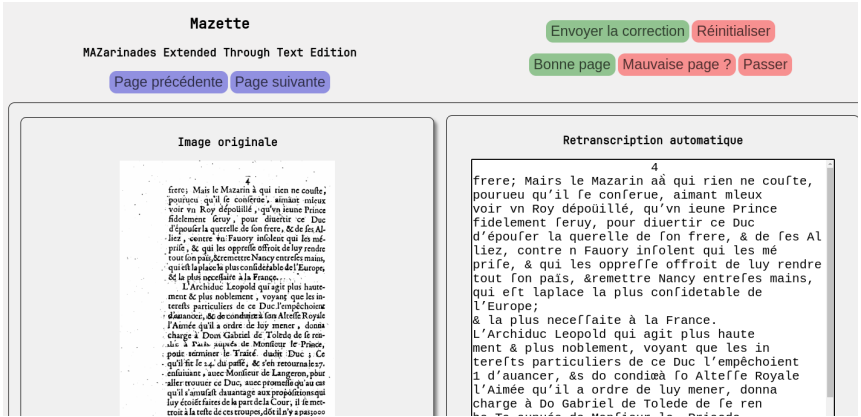


Figure: Capture d'écran de Mazette (`https://mazette.marceau-h.fr/`)

Mazette - Objectif

2 idées principales:

- Faire corriger des pages dont on connaît la correction afin d'évaluer la 'qualité' de l'annotateur, à la manière du reCATPCHA (von Ahn et al., 2008)
- Fusionner les corrections d'une même page

Version réduite du problème initial dans laquelle nous avons une source commune à chaque version d'un même document et pouvons calculer un score de confiance.

Merge without conflict ? Stratégies de fusion

Stratégies de fusion

2 approches se recoupant:

Méthode ROVER (Fiscus, 1997) Issue de l'ASR, le but est de faire un reconnaisseur *composite*, plusieurs outils pour produire une meilleure version

diff (Khanna et al., 2007) Méthode utilisée pour le versionnage de fichiers, reposant sur l'algorithme des *longest common substrings*

ROVER

Recognizer Output Voting Error Reduction (ROVER)

2 étapes:

Alignement En deux sous-étapes :

- Un réseau de transition par texte avec une transition par *mot*
- Alignement et fusion récursifs des transitions dans un même réseau

Vote En fonction de deux critères :

- Probabilité d'apparition du mot
- Score de confiance dans l'annotateur

ROVER: figures I

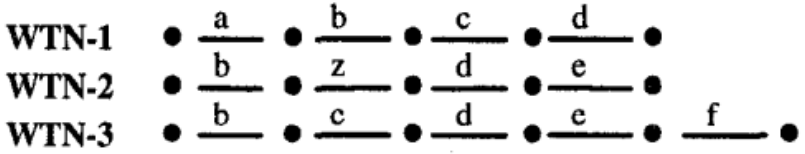


Figure: Réseaux initiaux des versions (Fiscus, 1997)

ROVER: figures II

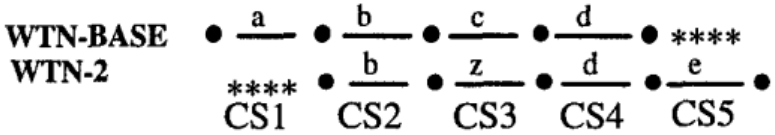


Figure: Alignement des réseaux 1 et 2, le réseau 1 devenant la ‘base‘ (Fiscus, 1997)

ROVER: figures III

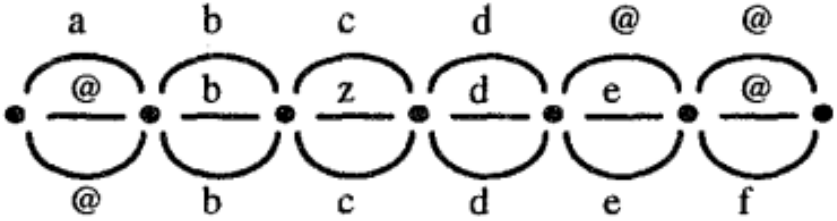
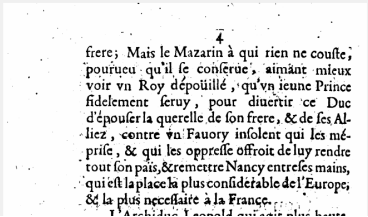


Figure: Graphe final, obtenu par fusion des 3 graphes initiaux (Fiscus, 1997)

ROVER: exemple I

Image originale

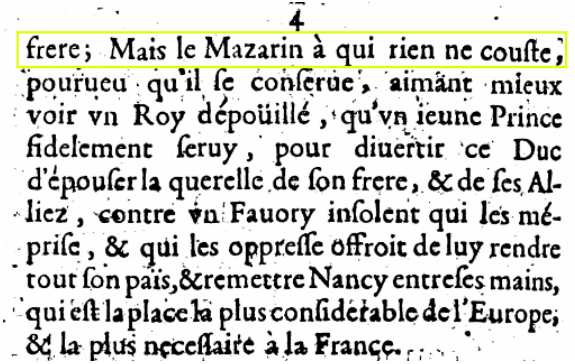


Retranscription automatique

```
4
frere; Mairs le Mazarin aà qui rien ne couſte,
pourueu qu'il ſe conſerue, aimant mleux
voir vn Roy dépoüillé, qu'vn ieune Prince
fidelement ſeruy, pour diuertir ce Duc
d'épouſer la querelle de ſon frere, & de ſes Al
liez, contre n Fauory inſolent qui les mé
priſe, & qui les oppreſſe offroit de luy rendre
tout ſon païs, &remettre Nancy entreſes mains,
qui eſt laplace la plus conſidetable de l'Europe;
& la plus neceſſaite à la France.
```

Figure: Exemple: capture d'écran sur le site

ROVER: exemple II

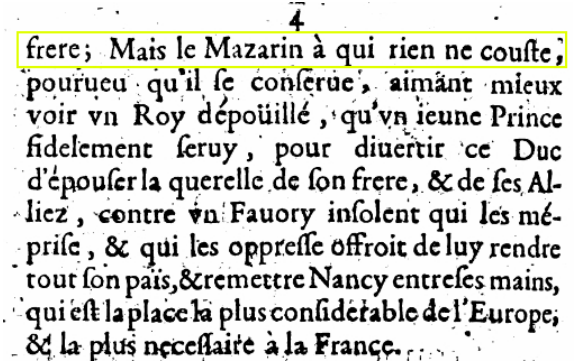


4

frere; Mais le Mazarin à qui rien ne couſte,
pourueu qu'il ſe conſerue, aimant mieux
voir vn Roy dépoüillé, qu'vn ieune Prince
fidelement ſeruy, pour diuertir ce Duc
d'épouſer la querelle de ſon frere, & de ſes Al-
liez, contre vn Fauory inſolent qui les mé-
priſe, & qui les oppreſſe offroit de luy rendre
tout ſon païs, & remettre Nancy entre ſes mains,
qui eſt la place la plus conſiderable de l'Europe,
& la plus neceſſaire à la France.

(a) Partie étudiée (ligne 2)

ROVER: exemple II



4

frere; Mais le Mazarin à qui rien ne couſte,
pourueu qu'il ſe conſerue, aimant mieux
voir vn Roy dépoüillé, qu'vn ieune Prince
fidelement ſeruy, pour diuertir ce Duc
d'épouſer la querelle de ſon frere, & de ſes Al-
liez, contre vn Fauory inſolent qui les mé-
priſe, & qui les oppreſſe offroit de luy rendre
tout ſon païs, & remettre Nancy entre ſes mains,
qui eſt la place la plus conſiderable de l'Europe,
& la plus neceſſaire à la France.

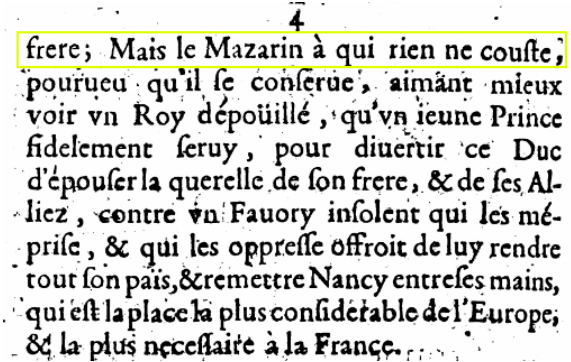
(a) Partie étudiée (ligne 2)

Retranscription automatique - confiance 10%
frere; Mairs le Mazarin aà qui rien ne couſte,

Annotateur 1 - confiance 50%
frere; Mairs le Mazarin à qui rien ne couſte,

Annotateur 2 - confiance 90%
frere; Mais le Mazarin à qui rien ne couſte,

ROVER: exemple II



4

frere; Mais le Mazarin à qui rien ne couſte,
pourueu qu'il ſe conſerue, aimant mieux
voir vn Roy dépoüillé, qu'vn ieune Prince
fidelement ſeruy, pour diuertir ce Duc
d'épouſer la querelle de ſon frere, & de ſes Al-
liez, contre vn Fauory inſolent qui les mé-
priſe, & qui les oppreſſe offroit de luy rendre
tout ſon païs, & remettre Nancy entre ſes mains,
qui eſt la place la plus conſiderable de l'Europe;
& la plus neceſſaire à la France.

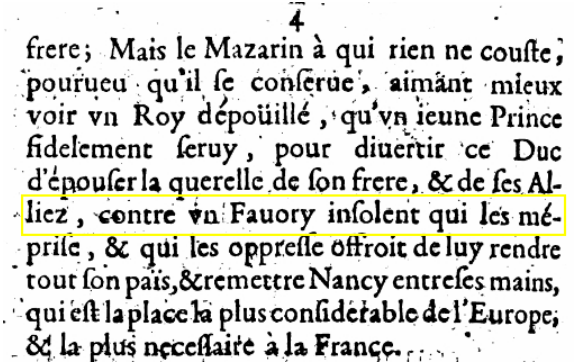
(a) Partie étudiée (ligne 2)

```
Retranscription automatique - confiance 10%
frere; Mairs le Mazarin aà qui rien ne couſte,

Annotateur 1 - confiance 50%
frere; Mairs le Mazarin à qui rien ne couſte,

Annotateur 2 - confiance 90%
frere; Mais le Mazarin à qui rien ne couſte,

r → 60
ø → 90

a → 10
ø → 140
```

(b) Fusion des différences

Figure: Exemple 1 : Fusion réussie

ROVER: exemple III

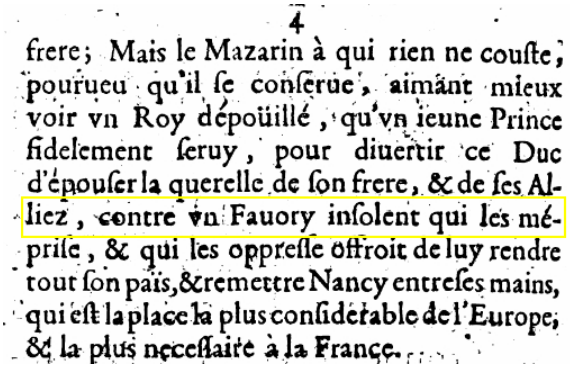


4

frere; Mais le Mazarin à qui rien ne couſte,
pourueu qu'il ſe conſerue, aimant mieux
voir vn Roy dépoüillé, qu'vn ieune Prince
fidelement ſeruy, pour diuertir ce Duc
d'épouſer la querelle de ſon frere, & de ſes Al-
liez, contre vn Fauory inſolent qui les mé-
priſe, & qui les oppreſſe offroit de luy rendre
tout ſon païs, & remettre Nancy entre ſes mains,
qui eſt la place la plus conſiderable de l'Europe,
& la plus neceſſaire à la France.

(a) Partie étudiée (ligne 7)

ROVER: exemple III



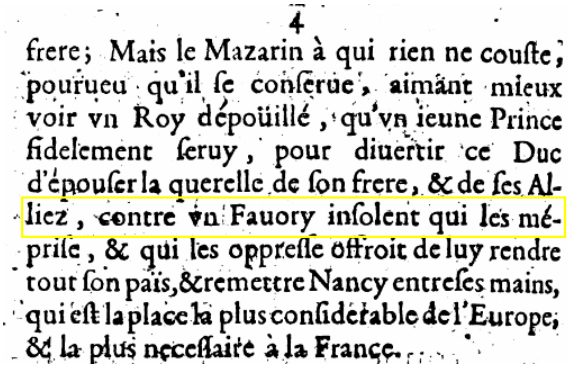
4

frere; Mais le Mazarin à qui rien ne couſte, pourueu qu'il ſe conſerue, aimant mieux voir vn Roy dépoüillé, qu'vn ieune Prince fidelement ſeruy, pour diuertir ce Duc d'épouſer la querelle de ſon frere, & de ſes Alliez, contre vn Fauory inſolent qui les mépriſe, & qui les oppreſſe offroit de luy rendre tout ſon païs, & remettre Nancy entre ſes mains, qui eſt la place la plus conſiderable de l'Europe, & la plus neceſſaire à la France.

(a) Partie étudiée (ligne 7)

```
Retranscription automatique - confiance 10%
liez, contre n Fauory inſolent qui les mé

Annotateur 1 - confiance 50%
liez, contre vn Fauory inſolent qui les mé

Annotateur 2 - confiance 50%
liez, contre un Fauory inſolent qui les mé
```

ROVER: exemple III



4

frere; Mais le Mazarin à qui rien ne couſte, pourueu qu'il ſe conſerue, aimant mieux voir vn Roy dépoüillé, qu'vn ieune Prince fidelement ſeruy, pour diuertir ce Duc d'épouſer la querelle de ſon frere, & de ſes Alliez, contre vn Fauory inſolent qui les mépriſe, & qui les oppreſſe offroit de luy rendre tout ſon païs, & remettre Nancy entre ſes mains, qui eſt la place la plus conſiderable de l'Europe; & la plus neceſſaire à la France.

(a) Partie étudiée (ligne 7)

```
Retranscription automatique - confiance 10%
liez, contre n Fauory inſolent qui les mé

Annotateur 1 - confiance 50%
liez, contre vn Fauory inſolent qui les mé

Annotateur 2 - confiance 50%
liez, contre un Fauory inſolent qui les mé

ø → 10
v → 50
u → 50
Pas de décision !
```

(b) Fusion des différences

Figure: Exemple 2 : Conflit

Exemple III, Schéma I

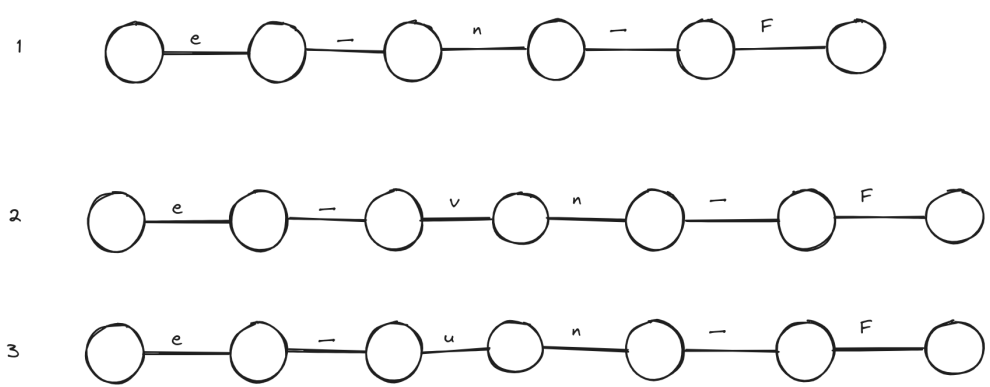


Figure: Schéma 1: création des 3 graphes, seulement sur le segment contenant la différence ici

Exemple III, Schéma II

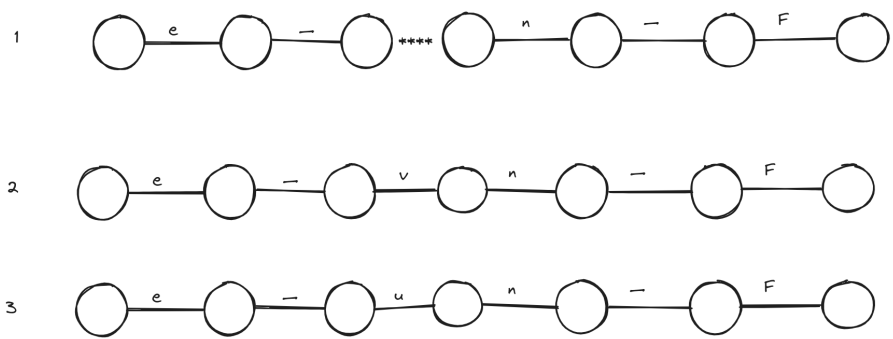


Figure: Schéma 2: alignement des graphes, transition absente représentée par des étoiles

Exemple III, Schéma III

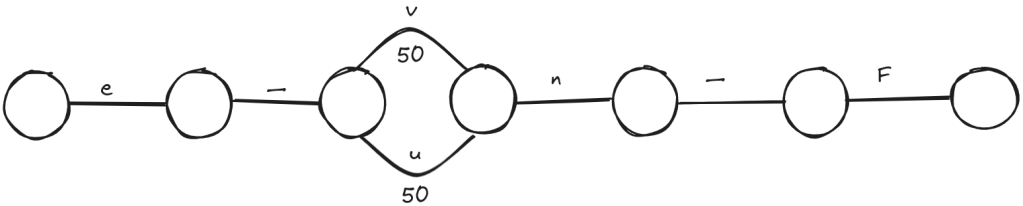


Figure: Schéma 3: fusion des graphes, poids représentés en dessous des transitions

diff (Khanna et al., 2007)

‘diff‘ calcule les changements entre deux fichiers, supposés proches.
Implémentation ‘diff3‘ utilisée par *GitHub* pour résoudre les changements faits à partir d’une même version.

diff (Khanna et al., 2007)

‘diff‘ calcule les changements entre deux fichiers, supposés proches.
Implémentation ‘diff3‘ utilisée par *GitHub* pour résoudre les changements faits à partir d’une même version.

Dans le cas d’un *3-way merge*, on a alors :

- Fusion des différences à partir d’un ancêtre commun
- Les parties identiques ne changent pas
- Les parties modifiées dans une seule version sont ajoutées

How to diff

2 étapes:

Alignement Calcul des *Longest Common Subsequences* (LCS), suites de texte identiques les plus longues et non contiguës

Agrégation des modifications Toutes les parties en dehors des LCS et non conflictuelles sont intégrées, comme si elles avaient été toutes deux écrites dans le fichier d'origine

How to diff

Mais 3 problèmes :

- Non gestion des modifications entre version se chevauchant
- Ne compare que 2 versions
- Besoin d'un ancêtre commun
  - Ok pour correction, moins pour transcriptions et éditions multiples (sources différentes / plusieurs sources par document)

Différences entre deux versions

Ceci
est
un
texte
qui
sera
modifié

(a) Version mère

Ceci
est
le
texte
modifié

(b) Version fille

Figure: Différences entre deux versions

Différences entre trois versions I

```
Ceci
est
un
texte
qui
sera
modifié
```

(a) Version mère

```
Ceci
est
le
texte
modifié
```

(b) Version fille 1

```
Ceci
est
un
texte
modifié
depuis
la
même
source
```

(c) Version fille 2

```
Ceci
est
le
texte
modifié
depuis
la
même
source
```

(d) Résultat de la fusion

Figure: Différences entre trois versions, fusion réussie

Différences entre trois versions II

```
Ceci
est
un
texte
qui
sera
modifié
```

(a) Version mère

```
Ceci
est
le
texte
modifié
```

(b) Version fille 1

```
Cela
est
mon
texte
modifié
```

(c) Version fille 2

Figure: Différences entre trois versions, conflit

Solution ?

Recursive 3-way merge

Fusion à partir d'un ancêtre virtuel, en l'absence d'un ancêtre commun unique

Lorsqu'une version est issue de plusieurs autres versions

Mais utilisable qu'en connaissant l'arbre des versions

Donc toujours besoin d'au moins un ancêtre commun.

Et non gestion des parties identiques modifiées

Et maintenant ?

Fusion.. Des fusions ?

Recoupements entre les mesures, alignement de paires de textes puis fusion
Différences clés:

- ROVER
  - \+ Fusion par système de vote, prend toujours une décision
  - \+ Fusion sans ancêtre commun

  Plus adapté à la fusion de versions sans ancêtre commun
- diff
  - \+ Fusion selon les modifications
  - \- Pas de décision en cas de conflit

  Plus adapté à la fusion de corrections de retranscriptions

Faudrait-il choisir la fusion selon la tâche ? Essayer de mélanger les idées ?

Et après ?

- Comment donner un poids aux versions ? Quels critères?
  - ancienneté
  - exhaustivité ou non de la version
  - confiance en la source
    - Score unique
    - Score selon l'entrée
    - Score selon la sortie
- Est-il possible de reconstituer ou de déterminer un arbre des versions existantes, avec des ancêtres manquants ?
- Autres méthodes de fusion ?

References I

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354.

Khanna, S., Kunal, K., and Pierce, B. C. (2007). A formal investigation of diff3. In *FSTTCS 2007: Foundations of Software Technology and Theoretical Computer Science: 27th International Conference, New Delhi, India, December 12-14, 2007. Proceedings 27*, pages 485–496. Springer.

von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468.