Utilizing Web Service Technology to Create Danish Arabic Language Resources

Mossab Al-Hunaity

Centre for Language Technology
University of Copenhagen

Agenda

- Research Problem
- Model Introduction
- Discussion

Research Problem

- Language resources LR are a major challenge for modern SMT applications for language with limited common LR like the case of Danish Arabic language pair.

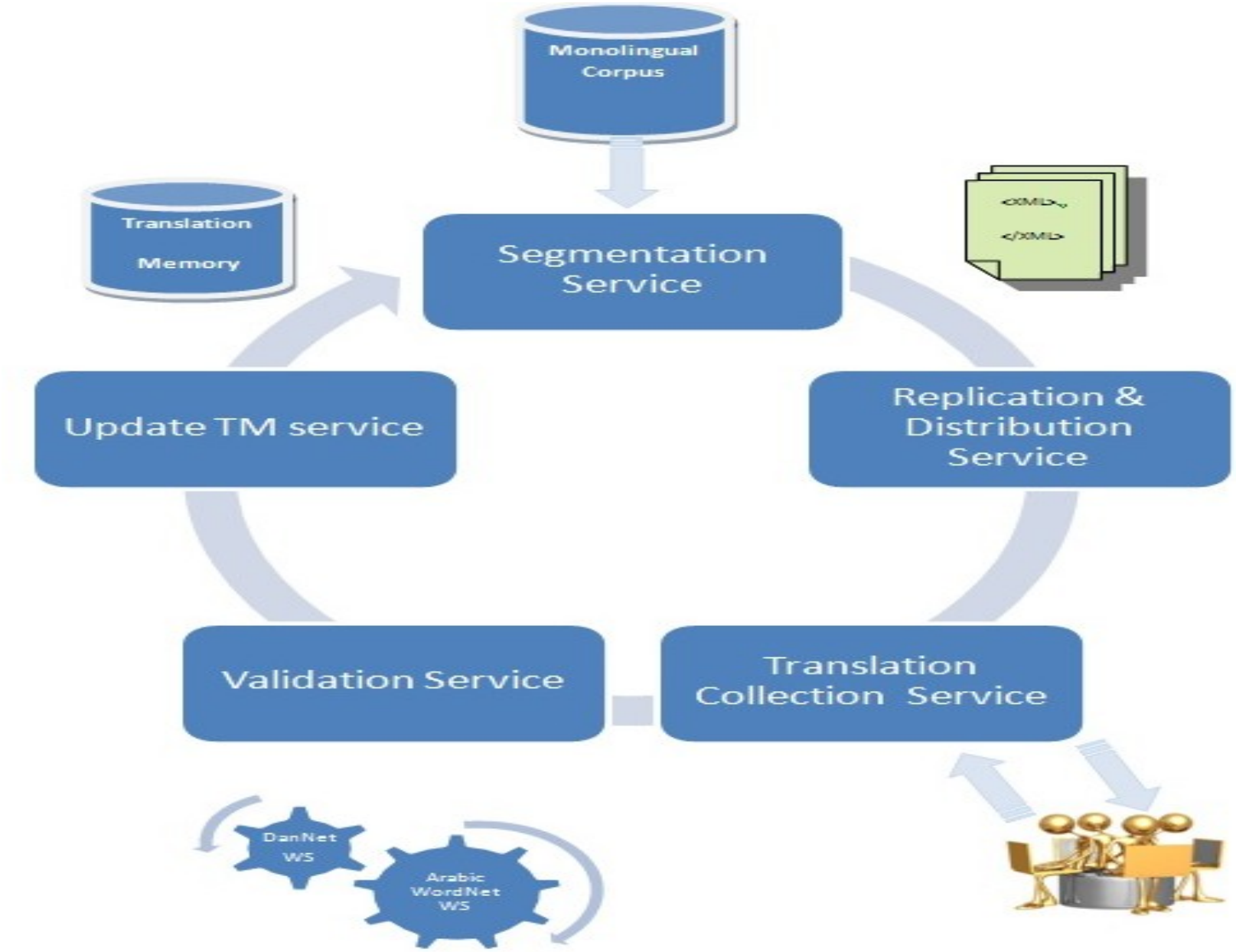
- Solution might be in the: Pivot approach
  - Translation quality is less than direct approaches

- We propose a new model that utilizes the web service technology to create a bilingual text resource out of a monolingual corpus.

Model Introduction

1. Segmentation Service.
2. Distribution and Replication Service.
3. Translation collection Service.
4. Validation Service.
5. Update translation memory Service.

Model Introduction Cont.



Segmentation Service 1

- The service will process the monolingual corpus and compile it into a group of small XML files

```
<?xml version="1.0" encoding="UTF-8"?>
<SRCSET setid="Climate_Change_Summit" srclang="DA">
    <DOC docid="1" genre = "text" >
      <seg id="1.1">
          de fire vigtige punkter, der bør rummes i en aftale i
          København
      </seg>
      <seg id="1.2">
          Hvor meget er industrilandene villige til at reducere deres
          udledning af drivhusgasser?
      </seg>
      <seg id="1.3">
          Hvor meget er toneangivende udviklingslande
          som Kina og Indien villige til at gøre for at begrænse
          stigningen i deres udledning?
    </seg>
    <seg id="1.4">
          Hvis København kan levere varen på de fire
          punkter, vil jeg være glad," siger Yvo de Boer.
    </seg>
  </DOC>
</SRCSET>
```

Segmentation Service 2

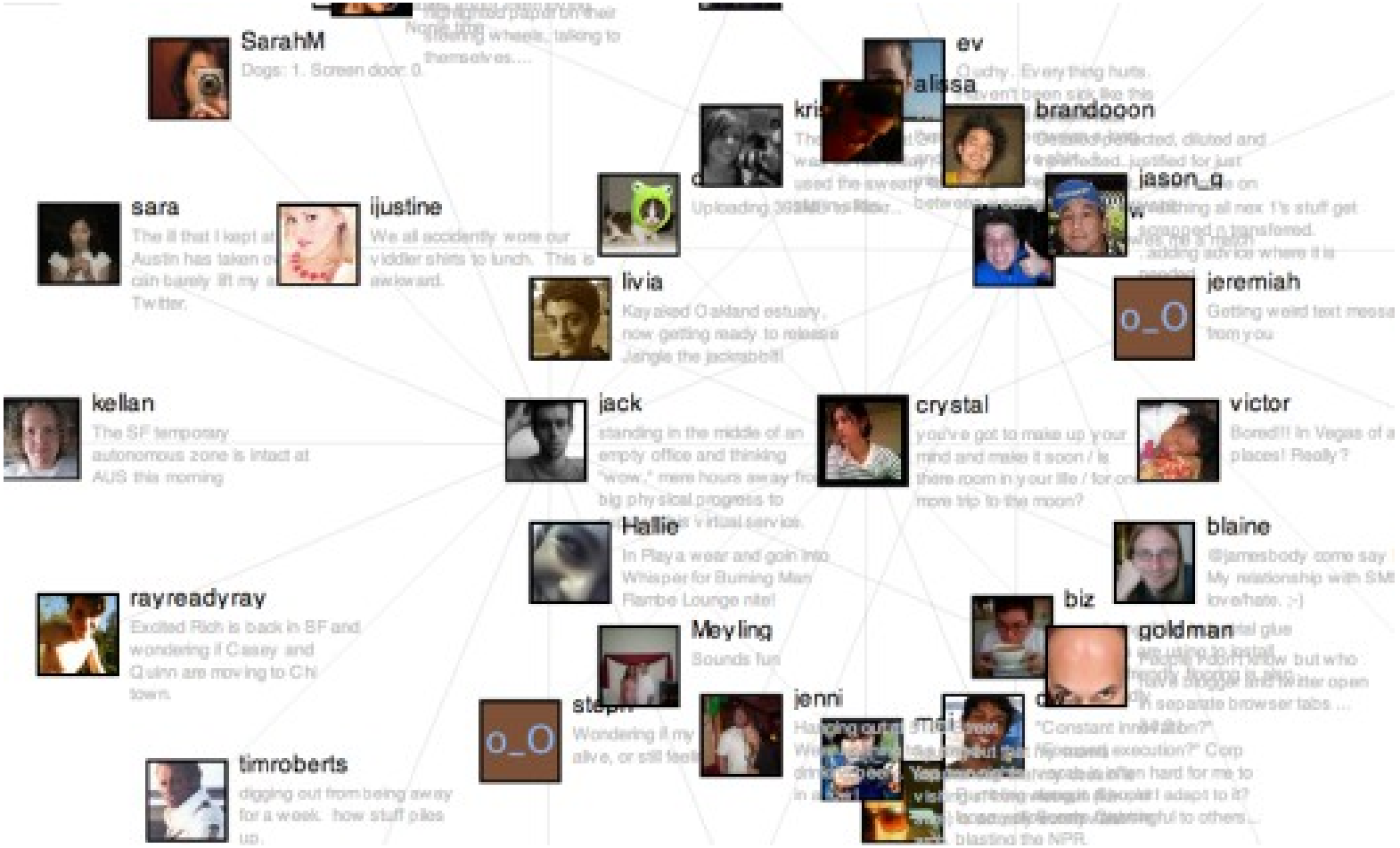
- Segments produces from a corpus document, ready to be sent to net work users

```
<?xml version="1.0" encoding="UTF-8"?>
<SRCSET setid="Climate_Change_Summit" srclang="DA">
      <DOC docid="1" genre = "text" >
         <seg id="1.1">
             de fire vigtige punkter, der bør rummes i en aftale i
             København
         </seg>
   </DOC>
</SRCSET>
```

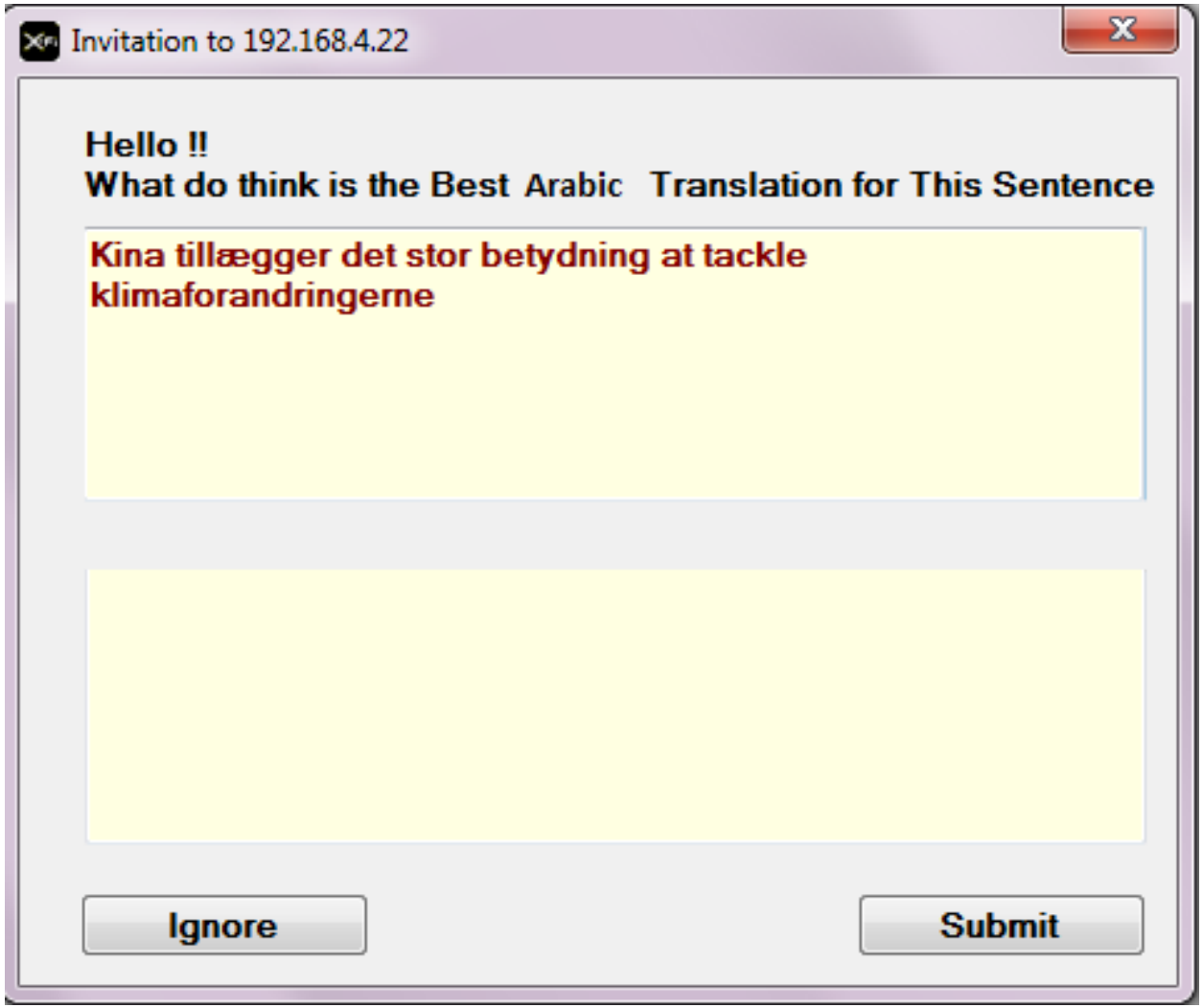
```
<?xml version="1.0" encoding="UTF-8"?>
<SRCSET setid="Climate_Change_Summit" srclang="DA">
      <DOC docid="1" genre = "text" >
         <seg id="1.2">
             Hvor meget er industrilandene villige til at reducere deres
              udledning af drivhusgasser?
         </seg>
   </DOC>
</SRCSET>
```

Replication and Distribution Service

- This service receives segments files produced by the segmentation service and distribute it to network users.



Replication and Distribution Service



Replication and Distribution Service

Replication and Distribution Service

Translation Collection

- Translation Collection service will collect translation from network users who agreed to respond to the translation request

```
<?xml version="1.0" encoding="UTF-8"?>
<SRCSET setid="Climate_Change_Summit" srclang="DA">
      <DOC docid="1" genre = "text" >

      <trans id="1" seg ="1.1"    user="192.168.4.22">
            المناخ لتغير  التصدي على  كبيرة أهمية  تعلق الصين
      </trans>

      <trans id="2" seg ="1.1"    user="192.168.8.36">
                  المناخ  بموضوع  تهتم الصين
      </trans>

      <trans id="3" seg ="1.1"    user="192.168.8.64">
                  اسيوية دولة الصين
      </trans>


   </DOC>
</SRCSET>
```

Translation validation service

1. Automatic Validation
   - Length Validation
   - Similarity Matching
   - Accuracy & Fitness Evaluation

1. Human Validation and Evaluation.

Length Validation

- We compare the length of the original sentence (S) to the length of the translated sentence (D) .

$$r = \frac{L(S)}{L(D)}$$

- The service accept a sentence D as a possible translation for sentence S if r > 0.75

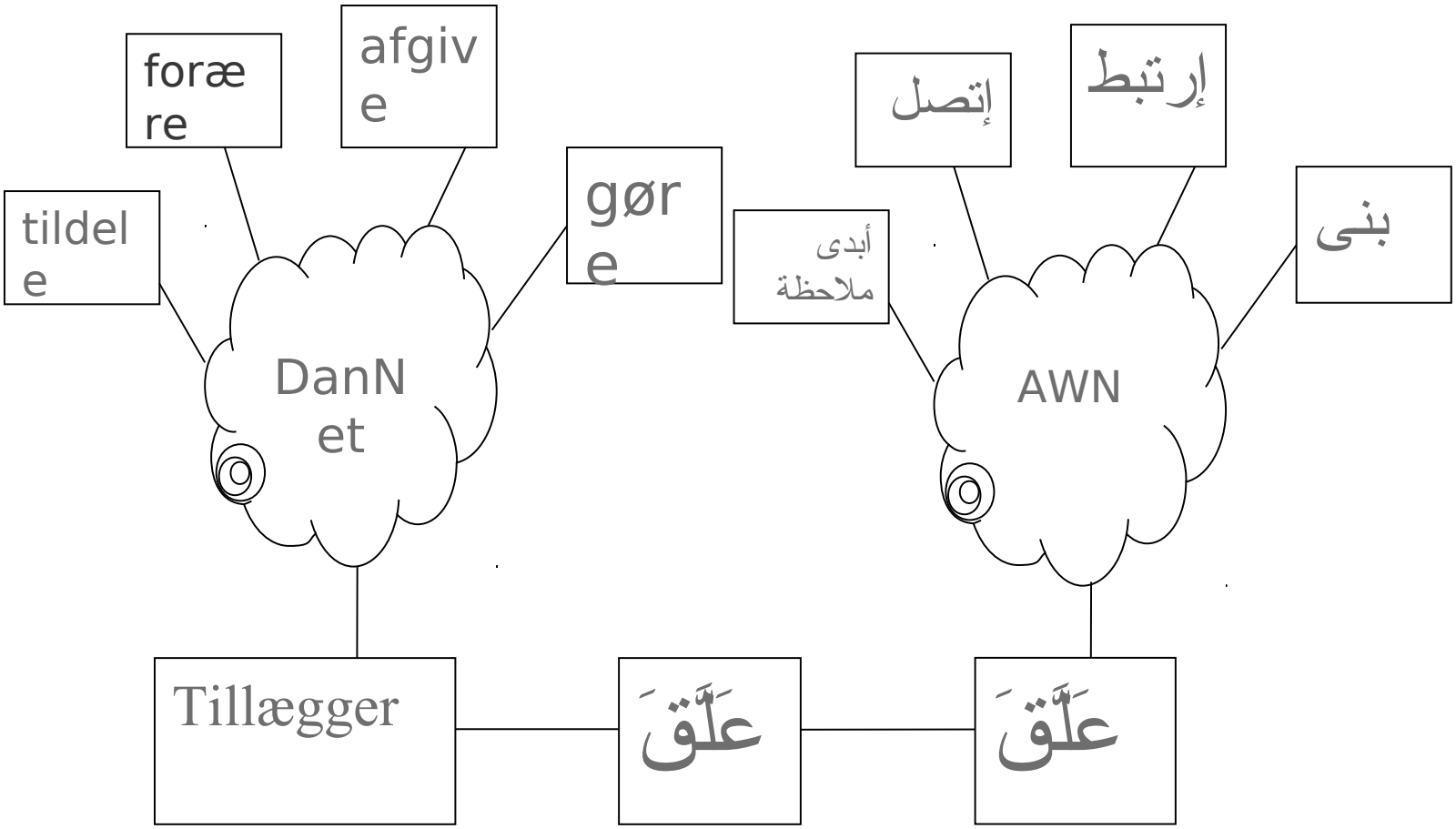
Similarity Matching

- Now for each sentences pair (S, D) we compare their similarities

S / D	Kina	tillægger	det	stor	betydning	at	handler	klimaændringerne	match (Si,Dj)
الصـــين	1	0	0	0	0	0	0	0	1
تعلـــق	0	1	0	0	0	0	0	0	1
أهمية	0	0	0	0	1	0	0	0	1
كبـــيرة	0	0	0	1	0	0	0	0	1
علـى	0	0	0	0	0	1	0	0	1
التصـــدي	0	0	0	0	0	0	1	0	1
لتغـــير	0	0	0	0	0	0	0	1	1
المنـاخ	0	0	0	0	0	0	0	1	1

$$\text{Sim (S,D)} = \frac{\sum_{J=1}^{n} \sum_{i=1}^{m} \text{match (Si,Dj)}}{\max(L(S), L(D))}$$

AWN and DanNet



Translation Accuracy & Fitness

Accuracy

الصين	تعلق	أهـمية	كبيرة	على	التـــصدي	لتغير	المـــناخ	∑
1/1	1/11	1/2	1/7	1/1	1/4	1/10	1/1	4.08

$$Acc(D) = \frac{\sum_{j=1}^{m} \left(\frac{1}{AWN(D(j))} \right)}{m}$$

Fitness

$$Fitness\ (S, D) = Sim(S, D) * Acc\ (D)$$

Heuristics
- **Accuracy > 0.4**
- **Fitness > 0.3**

Will represent acceptable translation

Human Evaluation

- System validation indicates whether the destination is a **good candidate translation** for the source sentence or not.
- It doesn't mean that the sentence is accepted
- Human evaluation is needed to accept the translation.

Update the Translation memory

```
<?xml version="1.0" encoding="UTF-8"?>
   <SRCSET setid="Climate_Change_Summit" srclang="DA", dstlang="AR">
      <DOC docid="1" genre = "text" >

      <trans id="1" seg ="1.1" srcdoc ="1"    >
            Kina tillægger det stor betydning at tackle klimaforandringerne
      </trans>

      <trans id="1" seg ="1.1"    user="192.168.4.22">
                                  المناخ لتغير  التصدي على  كبيرة أهمية  تعلق الصين
      </trans>

     </DOC>
   </SRCSET>
```

Questions