Metric	Spearman's rho
TESLA-F	.94
TESLA-M	.93
meteor-next-*	.92
1-TERp	.90
BLEU-4-v13a-c	.89

Table 1: Selected system-level Spearman's rho correlation with the human judgment for the into-English task, as reported in WMT 2010.

Metric	Spearman's rho
TESLA-M	.93
meteor-next-rank	.82
1-TERp	.81
BLEU-4-v13a-c	.80
TESLA-F	.76

Table 2: Selected system-level Spearman's rho correlation with the human judgment for the out-of-English task, as reported in WMT 2010.

tags. While such tools are usually available even for languages other than English, it does make TESLA-M more troublesome to port to non-English languages.

TESLA-M did well in the WMT 2010 evaluation campaign. According to the system-level correlation with human judgments (Tables 1 and 2), it ranks top for the out-of-English task and very close to the top for the into-English task (Callison-Burch et al., 2010).

2.4 TESLA-F[3]

TESLA-F builds on top of TESLA-M. While word-level synonyms are handled in TESLA-M by examining WordNet synsets, no modeling of phrase-level synonyms is possible. TESLA-F attempts to remedy this shortcoming by exploiting a phrase table between the target language and another language, known as the pivot language.

Assume the target language is English and the pivot language is French, i.e., we are provided with an English-French phrase table. Let R and T be the reference and the translation candidate respectively, both in English. As an example,

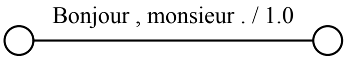

Figure 2: A degenerate confusion network in French. The phrase table maps *Good morning , sir .* to *Bonjour , monsieur .*

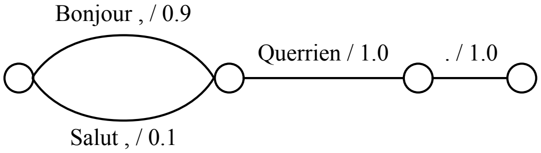

Figure 3: A confusion network in French. The phrase table maps *Hello ,* to *Bonjour ,* with $P = 0.9$ and to *Salut ,* with $P = 0.1$.

R: Good morning , sir .
T: Hello , Querrien .

TESLA-F first segments both R and T into phrases to maximize the probability of the sentences. For example, suppose both *Good morning , sir .* and *Hello ,* can be found in the English-French phrase table, and proper name *Querrien* is out-of-vocabulary, then a likely segmentation is:

R: ||| Good morning , sir . |||
T: ||| Hello , ||| Querrien ||| . |||

Each English phrase is then mapped to a bag of weighted French phrases using the phrase table, transforming the English sentences into confusion networks resembling Figures 2 and 3. French n-grams are extracted from these confusion network representations, known as pivot language n-grams. The bag of pivot language n-grams generated by R is then matched against that generated by T with the same linear programming formulation used in TESLA-M.

TESLA-F incorporates all the F-measures used in TESLA-M, with the addition of (1) the F-measures generated over the pivot language n-grams described above, and (2) the normalized language model score, defined as $\frac{1}{n} \log P$, where n is the length of the translation, and P the language model probability. Unlike BLEU and TESLA-M which rely on simple averages (geometric and arithmetic average respectively) to combine the component scores, TESLA-

[3] TESLA-F refers to the metric called TESLA in (Liu et al., 2010). To minimize confusion, in this work we call the metric TESLA-F and refer to the whole family of metrics as TESLA. F stands for *full.*